On the Convergence of the Metropolis Algorithm with Fixed-Order Updates for Multivariate Binary Probability Distributions

Kai Brügge
Dept. of Computer Science
University of Copenhagen
kai.brugge@gmail.com

Asja Fischer
Faculty of Mathematics
Ruhr University Bochum
asja.fischer@rub.de

Christian Igel
Dept. of Computer Science
University of Copenhagen
igel@di.ku.dk

Abstract

The Metropolis algorithm is arguably the most fundamental Markov chain Monte Carlo (MCMC) method. But the algorithm is not guaranteed to converge to the desired distribution in the case of multivariate binary distributions (e.g., Ising models or stochastic neural networks such as Boltzmann machines) if the variables (sites or neurons) are updated in a fixed order, a setting commonly used in practice. The reason is that the corresponding Markov chain may not be irreducible. We propose a modified Metropolis transition operator that behaves almost always identically to the standard Metropolis operator and prove that it ensures irreducibility and convergence to the limiting distribution in the multivariate binary case with fixed-order updates. The result provides an explanation for the behaviour of Metropolis MCMC in that setting and closes a long-standing theoretical gap. We experimentally studied the standard and modified Metropolis operator for models where they actually behave differently. If the standard algorithm also converges, the modified operator exhibits similar (if not better) performance in terms of convergence speed.

1 INTRODUCTION

Markov Chain Monte Carlo (MCMC) algorithms address the problem of sampling from a probability distribution p by constructing a Markov chain with stationary distribution equal to p. Recording a state of the chain after running it for some time replaces sampling from p. It has to be ensured that the Markov chain is ergodic, that is, converges to the stationary distribution irrespective of the starting state. The Metropolis algorithm (Metropolis et al., 1953) established the field of MCMC and is still widely used. It can be used as a building block for more advanced sampling algorithms such as Parallel Tempering (Geyer, 1991; Salakhutdinov, 2009; Fischer and Igel, 2015). However, for the basic scenario of multivariate binary distributions – as for example in Ising models or Boltzmann machines – and state-updates that consider these variables in a fixed order, it is well known that the Metropolis algorithm may not converge to the desired distribution (see, e.g., Friedberg and Cameron, 1970), because the Markov chain induced can be reducible.

Brügge et al. (2013) have suggested a slightly modified Metropolis algorithm for the sampling of restricted Boltzmann machines. We extend their work and prove that the modified algorithm induces ergodic Markov chains for all multivariate binary distributions (in the non-binary case the convergence problem does not occur, see Appendix A.1).

As a corollary we give a sufficient condition for the standard Metropolis algorithm to converge. For many classes of models, this condition is fulfilled almost surely. This theoretically justifies the use of the standard Metropolis algorithm with fixed-order updates.

In the next section, we state our main result, which is proven in Section 3. Section 4 provides numerical experiments before we conclude in Section 5.

2 MAIN RESULT

We consider the important case where p is an n-dimensional multivariate distribution with full support over a finite set Ω^n with binary Ω, covering Ising models (Ising, 1925; Brush, 1967) and stochastic neural networks such as restricted Boltzmann machines

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

(Smolensky, 1986; Hinton, 2002; Fischer and Igel, 2014). The transition probabilities of a (homogeneous) Markov chain with state-space Ω^n can be defined by a transition operator $\boldsymbol{T}$ representing the probabilities $\boldsymbol{T}(\boldsymbol{x} \to \boldsymbol{y})$ of going from state $\boldsymbol{x}$ to $\boldsymbol{y}$ in one step.

If p is a multivariate distribution, it is a common approach to consider a transition operator that is defined as a concatenation of operators of the form $\boldsymbol{T} = T_n \circ \dots \circ T_2 \circ T_1$, where operator T_i can only change the i-th variable (e.g., as in Gibbs sampling, Geman and Geman, 1984). We refer to the typical case of always applying the n operators in the same order as *fixed-order* updates. This corresponds to what Neal (1993) calls the *local* Metropolis algorithm.

For the standard *Metropolis algorithm*, the transition operator for the i-th variable is defined as:[1]

$$T_i(\boldsymbol{x} \to f_i(\boldsymbol{x})) = \begin{cases} 1 & \text{if } p(\boldsymbol{x}) \le p(f_i(\boldsymbol{x})) \\ \frac{p(f_i(\boldsymbol{x}))}{p(\boldsymbol{x})} & \text{if } p(\boldsymbol{x}) > p(f_i(\boldsymbol{x})) \end{cases} \quad (1)$$

for $\boldsymbol{x} = (x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)$ and $f_i(\boldsymbol{x}) = (x_1, \dots, x_{i-1}, \bar{x}_i, x_{i+1}, \dots, x_n)$, where $\bar{x}_i$ is the flipped value of x_i. We have $T_i(\boldsymbol{x} \to \boldsymbol{x}) = 1 - T_i(\boldsymbol{x} \to f_i(\boldsymbol{x}))$ and $T_i(\boldsymbol{x} \to \boldsymbol{y}) = 0$ for $\boldsymbol{y} \in \Omega^n \setminus \{\boldsymbol{x}, f_i(\boldsymbol{x})\}$. When using this operator, the Markov chain may not converge to p, because it may not be *irreducible*.

A Markov chain is irreducible if one can get from any state to any other in a finite number of transitions. Irreducibility is necessary for the chain to always converge to a unique stationary distribution. For fixed-order updates, the transition operators of the Metropolis algorithm do not necessarily lead to an irreducible Markov chain and the chain may not converge, potentially leading to a failure of the MCMC sampling algorithm. Examples are given in Section A in the appendix.

Updating the variables in random order simplifies the theoretical analysis and guarantees an irreducible Markov chain. In practice, though, a fixed order is usually preferred, because it can be implemented more efficiently. Additionally, fixed-order updates are usually assumed to lead to faster convergence (Roberts and Sahu, 1997; Levine, 2005; Andrieu, 2016; He et al., 2016) as they ensure that each variable is updated equally often and no variable is neglected for a longer time. He et al. (2016) shows that random scanning can indeed be worse by more than a logarithmic factor compared to fixed-order updates (in contrast to what is suggested by Diaconis, 2013, bottom p. 1299). Still, they can also lead to slower convergence or even a non-ergodic chain (Neal, 1993) (see also Appendix B). The importance of the update or scanning order for training machine learning models has long been realized, and several tailored algorithms have been proposed (see, e.g., He et al., 2016; Guo et al., 2018; Mitliagkas and Mackey, 2017).

For multivariate binary distributions, the Metropolis algorithm is similar to Gibbs sampling (Geman and Geman, 1984) – in fact Gibbs sampling can be seen as a Metropolis algorithm with a different proposal distribution or acceptance function (the Boltzmann acceptance function). A number of papers address the question of which of those two methods is preferable, see, e.g., Peskun (1973); Cunningham and Meijer (1974); Frigessi et al. (1993); Peskun (1981). Neal (1993) concludes: "The issues still remain unclear, though it appears that common opinion favors using the Metropolis acceptance function in most circumstances."

Brügge et al. (2013) proposed a slightly modified Metropolis operator for restricted Boltzmann machines, which we generalize to arbitrary binary distributions. The modified Metropolis transition operator only differs in the case when the current and proposed state have the same probability:

$$T_i(\boldsymbol{x} \to f_i(\boldsymbol{x})) = \begin{cases} 1 & \text{if } p(\boldsymbol{x}) < p(f_i(\boldsymbol{x})) \\ \frac{p(f_i(\boldsymbol{x}))}{p(\boldsymbol{x})} & \text{if } p(\boldsymbol{x}) > p(f_i(\boldsymbol{x})) \\ \frac{1}{2} & \text{otherwise} \end{cases} . \quad (2)$$

This modification ensures that p remains a stationary distribution, which is straight-forward to show by proving detailed balance. In the next section, we prove that for this operator it holds:

Theorem 1. *Let p be a distribution with full support over Ω^n for binary Ω and $n \ge 1$. The Markov chain induced by the modified Metropolis operator* (2) *and fixed-order updates is irreducible and aperiodic (and therefore ergodic).*

The derivation of this theorem relies on a novel proof strategy (construction of graphs where cycles correspond to contradictions and doing induction on these graphs). We use tools that, to our knowledge, have not been applied in that form to the analysis of MCMC algorithms. It would not have been possible, for example, to follow the approach by Brügge et al. (2013), which focuses on restricted Boltzmann machines and requires that the graph of the MRF is bipartite (the nodes are either *visible* or *hidden* units, and there are only connections between a visible and a hidden unit), while our proof is valid for all binary models (e.g., general Boltzmann machines).

[1] The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm. It requires the proposal function to be symmetric. For binary Ω we need not distinguish between proposal distribution and acceptance function, because flipping a state covers all proposal distributions.

From Theorem 1 follows our main result, a sufficient condition under which the standard Metropolis algorithm does not differ from the modified version and is therefore safe to use:

Corollary 1. *Let p be a distribution with full support over Ω^n for binary Ω and $n \geq 1$. If $p((x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)) \neq p((x_1,\dots,x_{i-1},\bar{x}_i,x_{i+2},\dots,x_n))$ for all $i = 1,\dots,n$, then the Markov chain induced by the standard Metropolis operator* (1) *and fixed-order updates is irreducible and aperiodic.*

For many use cases, for example spin glasses, where coupling strengths are drawn from a continuous distribution, or Boltzmann machines, where weights are initialized randomly, the condition in Corollary 1 holds almost always. Ising models with uniform coupling strength are an important case where the condition does not hold (see Appendix A.2). We suggest to use the modified Metropolis operator in cases where the standard operator would not lead to an irreducible chain. Section 4 demonstrates the performance of the new operator empirically.

3 PROOF OF MAIN RESULT

To prove Theorem 1, we use a basic theorem (e.g., see Billingsley, 1995, we refer to Hobert et al., 2007, for a proof):

Theorem 2. *A set $C \subseteq \Omega^n$ is* closed *given a Markov chain if for all $\boldsymbol{x} \in C : \sum_{\boldsymbol{y}\in C} \boldsymbol{T}(\boldsymbol{x} \rightarrow \boldsymbol{y}) = 1$ (i.e., once the chain enters C it cannot leave). A Markov chain is irreducible if the only closed subset is $C = \Omega^n$.*

We show irreducibility by proving that there exists no proper subset of Ω^n that cannot be left and then applying Theorem 2. The proof involves the following steps. We first establish a relation between possible transitions between states that differ only in one variable i and the probabilities of states resulting from flipping all variables with a higher (or lower) index than i including and excluding i. Then we show via contradiction that all singleton subsets of the state space cannot be closed. Then we show the same for sets with more than one element. To this end, we map each subset to a graph that contains cycles if the subset can be left and prove by induction that, for all proper subsets S, the graph $G(S)$ contains cycles and, therefore, the Markov chain is irreducible. Showing aperiodicity is straightforward.

Basic Lemma Let us denote the state $(\bar{x}_1,\dots,\bar{x}_n)$, where all sites (variables) are flipped to their opposite values, by $\overline{\boldsymbol{x}}$. We denote the state $(x_1,\dots,x_{i-1},\bar{x}_i,x_{i+1},\dots,x_n)$, where only the i-th site is flipped, by $f_i(\boldsymbol{x})$, the state where the first i sites are flipped by $f_{\leq i}(\boldsymbol{x}) = (\bar{x}_1,\dots,\bar{x}_i,x_{i+1},\dots x_n)$, and the state where the last i sites are flipped by $f_{\geq i}(\boldsymbol{x}) = (x_1,\dots,x_{i-1},\bar{x}_i,\dots,\bar{x}_n)$. We define the special boundary cases

$$f_{\leq 0}(\boldsymbol{x}) = f_{\geq n+1}(\boldsymbol{x}) = \boldsymbol{x} \ . \tag{3}$$

If a state $\boldsymbol{y}$ can be reached from a state $\boldsymbol{x}$ in an arbitrary number of steps, we write $\boldsymbol{x} \rightarrow \boldsymbol{y}$. If $\boldsymbol{y}$ cannot be reached from $\boldsymbol{x}$ in any numbers of steps, we write $\boldsymbol{x} \not\rightarrow \boldsymbol{y}$.

The following lemma establishes a relationship between properties of the stationary distribution p and impossible transitions of the Markov chain.

Lemma 1. *For all $i = 1,\dots,n$:*

$$\boldsymbol{x} \not\rightarrow f_i(\boldsymbol{x}) \Rightarrow p(f_{\leq i-1}(\boldsymbol{x})) < p(f_{\leq i}(\boldsymbol{x})) \tag{a}$$

$$\boldsymbol{x} \not\rightarrow f_i(\boldsymbol{x}) \Rightarrow p(f_{\geq i}(\boldsymbol{x})) < p(f_{\geq i+1}(\boldsymbol{x})) \tag{b}$$

Proof. For the proof of (a) we assume

$$p(f_{\leq i-1}(\boldsymbol{x})) \geq p(f_{\leq i}(\boldsymbol{x})) \tag{4}$$

and show that $\boldsymbol{x} \rightarrow f_i(\boldsymbol{x})$ follows. We do this by constructing a chain of valid transitions from $\boldsymbol{x}$ to $f_i(\boldsymbol{x})$. Note, that it is not sufficient to show that one can reach $f_i(\boldsymbol{x})$ by applying only part of the transition operators $T_i \circ \dots \circ T_2 \circ T_1$, for $i < n$, but one has also to show that it is possible to stay in this state until the completion of the full update step (i.e., under the remaining transition operators $T_n \circ \dots \circ T_{i+1}$).

Flipping a site is always possible, because $0 < p(f_{\leq i}(\boldsymbol{x}))/p(f_{\leq i-1}(\boldsymbol{x}))$, since we assume p having full support. Thus, for $i > 1$ we can transition from $\boldsymbol{x}$ to $f_{\leq i-1}(\boldsymbol{x})$ in the first $i-1$ partial steps. We can then stay in this state in the i-th partial transition step by assumption (4) (we need the assumption because staying in a state is only possible if it has larger or equal probability than the newly proposed state). The latter argument also proves $\boldsymbol{x} \rightarrow f_{\leq i-1}(\boldsymbol{x})$ for $i = 1$, which corresponds to staying in $\boldsymbol{x}$, see (3).

We continue to flip all the sites with the partial transition operators and end up in $f_{\geq i+1}(f_{\leq i-1}(\boldsymbol{x})) = \overline{f_i(\boldsymbol{x})}$, the state where all sites but the i-th site are flipped, at the end of the complete update step (i.e., after all n partial update steps). We can then transition to $f_i(\boldsymbol{x})$ by flipping all sites in another complete update step. Thus, we have created a path from $\boldsymbol{x}$ to $f_i(\boldsymbol{x})$ using two full update steps, which contradicts $\boldsymbol{x} \not\rightarrow f_i(\boldsymbol{x})$ in (a).

To prove (b), we analogously construct a path involving two full update steps by first flipping all sites and then flipping all sites but the i-th. □

The ability to stay in a certain state is the one point in the proof that works only for the modified Metropolis algorithm, but not the classical Metropolis algorithm. With the classical algorithm it is not possible to stay in the current state if $p(f_{\leq i-1}(\boldsymbol{x})) = p(f_{\leq i}(\boldsymbol{x}))$, while with the modified Metropolis operator it is. This also explains why the proof carries over to the classical Metropolis algorithm if there does not exist a $\boldsymbol{x} \in \Omega^n$ with $p(f_{\leq i-1}(\boldsymbol{x})) = p(f_{\leq i}(\boldsymbol{x}))$.

Singleton Subsets Cannot Be Closed Now we show that a singleton subset of the discrete state space cannot form a closed set:

Lemma 2. *Let $S = \{\boldsymbol{x}\} \subset \Omega^n$. It holds $\exists \boldsymbol{y} \in \Omega^n \setminus S : \boldsymbol{x} \to \boldsymbol{y}$.*

Proof. We prove the claim by showing that the negation

$$\forall \boldsymbol{y} \in \Omega^n \setminus S : \boldsymbol{x} \not\to \boldsymbol{y} \tag{5}$$

leads to a contradiction. Assuming (5), it follows that all states which differ from $\boldsymbol{x}$ in only one site cannot be reached from $\boldsymbol{x}$, that is:

$$\boldsymbol{x} \not\to f_i(\boldsymbol{x}) \text{ for } i = 1, \ldots, n \ . \tag{6}$$

Now, from Lemma 1 (a) follows that for $i = 1, \ldots, n$: $p(f_{\leq i-1}(\boldsymbol{x})) < p(f_{\leq i}(\boldsymbol{x}))$. From these inequalities we can construct the following sequence

$$p(\boldsymbol{x}) < p(f_{\leq 1}(\boldsymbol{x})) < .... < p(f_{\leq n-1}(\boldsymbol{x})) < p(\overline{\boldsymbol{x}}) \ .$$

From Lemma 1 (b) equivalently follows a second sequence

$$p(\overline{\boldsymbol{x}}) < p(f_{\geq 2}(\boldsymbol{x})) < \cdots < p(f_{\geq n}(\boldsymbol{x})) < p(\boldsymbol{x}) \ .$$

Together these two sequences of inequalities lead to the contradiction $p(\boldsymbol{x}) < p(\boldsymbol{x})$. □

Reduction to a Graph Problem Next, we generalize Lemma 2 to arbitrary subsets of Ω^n. That is, we want to prove that for all strict subsets $S \subset \Omega^n$, $\exists \boldsymbol{x} \in S$ and $\exists \boldsymbol{y} \in \Omega^n \setminus S$ such that $\boldsymbol{x} \to \boldsymbol{y}$ and thus S cannot be a closed set. The proof follows a similar line of thoughts as the proof of Lemma 2. We show that assuming the contrary, namely that there exists a $S \subset \Omega^n$ such that

$$\forall \boldsymbol{x} \in S : \forall \boldsymbol{y} \in \Omega^n \setminus S : \boldsymbol{x} \not\to \boldsymbol{y} \tag{7}$$

and therefore specifically

$$\forall \boldsymbol{x} \in S : \forall f_i(\boldsymbol{x}) \in \Omega^n \setminus S : \boldsymbol{x} \not\to f_i(\boldsymbol{x}) \ , \tag{8}$$

leads to a contradiction. Since reasoning about sequences of inequalities gets complicated when dealing with larger sets, we reduce the problem to a graph problem via mapping each subset S to a graph $G(S)$, such that each inequality resulting from applying Lemma 1 to statement (8) corresponds to an edge in the graph, and a contradiction to statement (8) arises if the graph contains cycles.

Lemma 3. *Let $S \subset \Omega^n$ and let $G(S) = (\Omega^n, E(S))$ be defined as the directed graph with edge set*

$$\begin{aligned} E(S) = \{ \big(f_{\leq i-1}(\boldsymbol{x}), f_{\leq i}(\boldsymbol{x})\big), \big(f_{\geq i}(\boldsymbol{x}), f_{\geq i+1}(\boldsymbol{x})\big) \mid \\ \boldsymbol{x} \in S \wedge f_i(\boldsymbol{x}) \in \Omega^n \setminus S\} \ . \end{aligned} \tag{9}$$

If $G(S)$ contains a cycle, then at least one state $f_i(\boldsymbol{x})$ outside S can be reached from some $\boldsymbol{x} \in S$, that is

$$\exists \boldsymbol{x} \in S : \exists f_i(\boldsymbol{x}) \in \Omega^n \setminus S : \boldsymbol{x} \to f_i(\boldsymbol{x}) \ . \tag{10}$$

Proof. Assume that for all $\boldsymbol{x} \in S$ we have $\forall f_i(\boldsymbol{x}) \in \Omega^n \setminus S : \boldsymbol{x} \not\to f_i(\boldsymbol{x})$. Then Lemma 1 states that $(\boldsymbol{x}, \boldsymbol{y}) \in E(S)$ implies the *strict* inequality $p(\boldsymbol{x}) < p(\boldsymbol{y})$. That is, if $G(S)$ contains a cycle, then the assumption cannot be fulfilled. □

To detect cycles, we make use of a basic theorem from graph theory. A directed graph is *balanced* if $\deg^+(v) = \deg^-(v)$ for every vertex v, where $\deg^+(v)$ and $\deg^-(v)$ are the numbers of outgoing and ingoing edges of v, respectively.

Theorem 3 (e.g., see Wahlström, 2018, p. 174)**.** *Let G be a balanced directed graph with at least one edge. Then there exists a directed cycle.*

Proof of Theorem 1. To prove **irreducibility**, we prove that for all $S \subset \Omega^n$ the graph $G(S)$ as defined in Lemma 3 contains a cycle. More specifically, we will apply Theorem 3 after showing that for all $S \subset \Omega^n$ the graph $G(S)$ has the property

$$\deg^+(v) = \deg^-(v) \ \ , v \in \Omega^n \ . \tag{11}$$

From the existence of a cycle for all $S \subset \Omega^n$, we know that no $S \subset \Omega^n$ can be a closed set by applying Lemma 3. Having shown that there is no proper subset of the state space that is a closed set, applying Theorem 2 yields that the Markov chain is irreducible.

We prove the property by induction. For the **base case**, we consider singleton subsets $\{\boldsymbol{x'}\}$ and the corresponding edge set:

$$\begin{aligned} E(\{\boldsymbol{x'}\}) = \{ \big(f_{\leq i-1}(\boldsymbol{x'}), f_{\leq i}(\boldsymbol{x'})\big), \\ \big(f_{\geq i}(\boldsymbol{x'}), f_{\geq i+1}(\boldsymbol{x'})\big) \mid i = 1, \ldots n\} \ . \end{aligned} \tag{12}$$

These edges form a cycle, which corresponds exactly to the contradiction arising from the sequence of inequalities discussed in the proof of Lemma 2. Each node of the cycle has exactly one incoming and one outgoing edge, and thus property (11) holds.

Let us now assume that $\forall S \subset \Omega^n$ with $|S| = k \geq 1$, for $G(S) = (\Omega^n, E(S))$ induced by (8) via Lemma 1 property (11) holds. In the **induction step** we now show that then for all $S' \subset \Omega^n$ with $|S'| = k+1$ property (11) also holds for $G(S') = (\Omega^n, E(S'))$.

For each S' it holds $S' = S \cup \{\boldsymbol{x}'\}$ for some $S \subset \Omega^n, |S| = k$ and $\boldsymbol{x}' \in \Omega^n \setminus S$. Given $G(S) = (\Omega^n, E(S))$ and $G(\{\boldsymbol{x}'\}) = (\Omega^n, E(\{\boldsymbol{x}'\}))$, the edge set $E(S')$ can be constructed as follows. Let $I = \{i \,|\, f_i(\boldsymbol{x}') \in S; i = 1, \dots, n\}$:

Case 1: Assume that $I = \emptyset$, that is, for all $i = 1, \dots, n$ it holds $f_i(\boldsymbol{x}') \notin S$, or equivalently $f_i(\boldsymbol{x}') \in \Omega^n \setminus S$, and therefore $f_i(\boldsymbol{x}') \in \Omega^n \setminus S'$. Directly from the definitions we have

$$\begin{aligned} E(S) = \{ \big(f_{\le i-1}(\boldsymbol{x}), f_{\le i}(\boldsymbol{x})\big), \big(f_{\ge i}(\boldsymbol{x}),& \\ f_{\ge i+1}(\boldsymbol{x})\big) \,|\, \boldsymbol{x} \in S \wedge f_i(\boldsymbol{x}) \in \Omega^n \setminus S\} =& \\ \{ \big(f_{\le i-1}(\boldsymbol{x}), f_{\le i}(\boldsymbol{x})\big), \big(f_{\le i-1}(\overline{\boldsymbol{x}}),& \\ f_{\le i}(\overline{\boldsymbol{x}})\big) \,|\, \boldsymbol{x} \in S \wedge f_i(\boldsymbol{x}) \in \Omega^n \setminus S\}& \end{aligned} \tag{13}$$

and

$$\begin{aligned} E(\{\boldsymbol{x}'\}) = \{ \big(f_{\le i-1}(\boldsymbol{x}'), f_{\le i}(\boldsymbol{x}')\big),& \\ \big(f_{\ge i}(\boldsymbol{x}'), f_{\ge i+1}(\boldsymbol{x}')\big) \,|\, i = 1, \dots n\} =& \\ \{ \big(f_{\le i-1}(\boldsymbol{x}'), f_{\le i}(\boldsymbol{x}')\big),& \\ \big(f_{\le i}(\overline{\boldsymbol{x}'}), f_{\le i+1}(\overline{\boldsymbol{x}'})\big) \,|\, i = 1, \dots n\} \ .& \end{aligned} \tag{14}$$

If $\overline{\boldsymbol{x}'} \in S$ then the edges in $E(\{\boldsymbol{x}'\})$ are already in $E(S)$ and $E(S \cup \{x'\}) = E(S)$. By assumption $\boldsymbol{x}' \notin S$. Thus, if $\overline{\boldsymbol{x}'} \notin S$ then $E(S)$ and $E(\{\boldsymbol{x}'\})$ are disjoint and

$$\begin{aligned} E(S') = \{ \big(f_{\le i-1}(\boldsymbol{x}), f_{\le i}(\boldsymbol{x})\big),& \\ \big(f_{\ge i}(\boldsymbol{x}), f_{\ge i+1}(\boldsymbol{x})\big) \,|\, \boldsymbol{x} \in S \cup \{\boldsymbol{x}'\} \wedge f_i(\boldsymbol{x}) \in \Omega^n \setminus S'\}& \\ = E(S) \cup E(\{\boldsymbol{x}'\}) \ .& \end{aligned} \tag{15}$$

In both cases $\deg^+(v) = \deg^-(v)$ for all vertices of $G(S')$.

Case 2: Let $I = \{i \,|\, f_i(\boldsymbol{x}') \in S; i = 1, \dots, n\} \neq \emptyset$. The set $E(S')$ can be constructed by removing edges from $E(S) \cup E(\{\boldsymbol{x}'\})$. This is illustrated in Figure 1. We need to remove the edges

$$\begin{aligned} R_i^S = \{ \big(f_{\le i-1}(\boldsymbol{x}), f_{\le i}(\boldsymbol{x})\big),& \\ \big(f_{\ge i}(\boldsymbol{x}), f_{\ge i+1}(\boldsymbol{x})\big) \,|\, \boldsymbol{x} \in S \wedge f_i(\boldsymbol{x}) = \boldsymbol{x}'\} \subseteq E(S)& \end{aligned} \tag{16}$$

for $i \in I$, because $\boldsymbol{x}'$ is in the complement of S, but not in the complement of S' and thus $R_i^S \nsubseteq E(S')$,[2] as well as the edges from

$$\begin{aligned} R_i^{\{\boldsymbol{x}'\}} = \{ \big(f_{\le i-1}(\boldsymbol{x}'),& \\ f_{\le i}(\boldsymbol{x}')\big), \big(f_{\ge i}(\boldsymbol{x}'), f_{\ge i+1}(\boldsymbol{x}')\big)\} \subseteq E(\{\boldsymbol{x}'\})& \end{aligned} \tag{17}$$

for each $i \in I$, because $f_i(\boldsymbol{x}') \notin \Omega^n \setminus S'$ implies $R_i^{\{\boldsymbol{x}'\}} \nsubseteq E(S')$. We can rewrite

$$\begin{aligned} R_i^S = \{ \big(f_{\le i-1}(f_i(\boldsymbol{x}')), f_{\le i}(f_i(\boldsymbol{x}'))\big),& \\ \big(f_{\ge i}(f_i(\boldsymbol{x}')), f_{\ge i+1}(f_i(\boldsymbol{x}'))\big)\}& \\ = \{ (f_{\le i}(\boldsymbol{x}'), f_{\le i-1}(\boldsymbol{x}')), (f_{\ge i+1}(\boldsymbol{x}'), f_{\ge i}(\boldsymbol{x}'))\} \ .& \end{aligned} \tag{18}$$

Comparing (17) to (18) shows that for each edge from $E(\{\boldsymbol{x}'\})$ not included in $E(S')$ we find a reverse edge from $E(S)$ not included in $E(S')$. From this it follows that $\deg^+(v) = \deg^-(v)$ for all vertices of $G(S')$. Thus, for all proper subsets $S \subset \Omega^n$ the graph $G(s)$ contains a cycle because of Theorem 3. This implies that S cannot be a closed set by Lemma 3. Thus, no proper subset of the state space is a closed set and the Markov chain is irreducible by Theorem 2.

To prove **aperiodicity** of an irreducible Markov chain we only need to identify one aperiodic state. There is one state $\boldsymbol{x} \in \Omega^n$ with $p(\boldsymbol{x}) \ge p(\boldsymbol{y})$ for all $\boldsymbol{y} \in \Omega^n$. By definition of the modified Metropolis operator (2) (and also the standard operator (1)), there is always the possibility that after reaching $\boldsymbol{x}$ the chain also stays in the state $\boldsymbol{x}$. Thus, the state $\boldsymbol{x}$ has period one and the Markov chain is aperiodic. □

4 EXPERIMENTS

Our main result is theoretical and contributes to the basic understanding of the general Metropolis algorithm, which has many applications in machine learning. Corollary 1 guarantees the irreducibility of the Metropolis chain for all binary Gibbs models for which the modified and the vanilla Metropolis algorithm are identical. Still, it is interesting to look at scenarios where the modified Metropolis algorithm differs from the original. We performed experiments on Ising models (with uniform coupling strength and no external field) (Ising, 1925; Brush, 1967), which belong to the class of models for which the condition in Corollary 1 does not hold. We selected Ising models because they are conceptually simple and have been used to demonstrate convergence problems of the Metropolis algorithm before. Most importantly for Ising models the modified and standard Metropolis operator actually differ, while for other popular models like Boltzmann machines or spin glasses our contribution is the proof of irreducibility which transfers to the standard operator according to Corollary 1. Numerous MCMC methods have been proposed since the Metropolis algorithm was invented, for example methods that flip clusters of variables (Swendsen and Wang, 1987; Wolff, 1989) developed to speed up the convergence when sampling Ising models. We do not claim the superiority of the

[2] For $i \notin I$ we have $R_i^S = \emptyset$, as $f_i(\boldsymbol{x}) = \boldsymbol{x}'$ would imply $\boldsymbol{x} = f_i(\boldsymbol{x}')$.

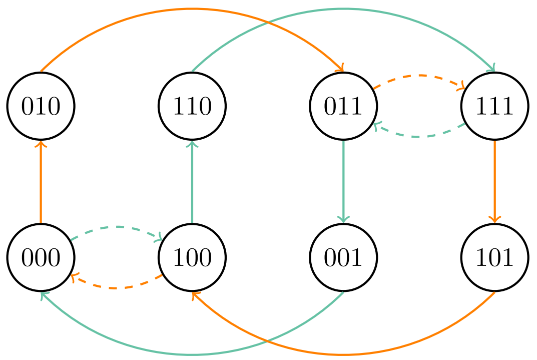

Figure 1: Example for the induction step in the proof of Theorem 1. We consider the state space $\{0,1\}^3$, the set $S = \{(0,0,0)\}$, and $\boldsymbol{x}' = (1,0,0)$. Both S and $\{\boldsymbol{x}'\}$ are singleton sets, so their graphs $G(S)$ and $G(\{\boldsymbol{x}'\})$ form one cycle each, corresponding to the cycle of inequalities described in the proof of Lemma 2. For each edge dropped from $G(S)$ there is an edge dropped from $G(\{\boldsymbol{x}'\})$ in the opposite direction. This means that for each vertex v of $G(S')$, the property $\deg^+(v) = \deg^-(v)$ still holds and the remaining edges form disjunct cycles, in this case two cycles. The figure shows $G(\{(0,0,0)\})$ in green, $G(\{(1,0,0)\})$ in orange, $G(\{(0,0,0),(1,0,0)\})$ with solid lines and the edges that are removed with dashed lines.

original nor the modified Metropolis algorithm, and the following experiments were conducted to illustrate the issues addressed in our theoretical work – they are not meant to be a performance comparison of MCMC algorithms, which would require a different experimental setup and a comparison of numerous algorithms (e.g., Swendsen and Wang, 1987; Wolff, 1989; He et al., 2016; Mitliagkas and Mackey, 2017; Guo et al., 2018), most of which would converge faster than Metropolis.

We applied the original Metropolis algorithm, the modified Metropolis algorithm as well as Gibbs sampling to small two-dimensional ferromagnetic Ising models (with no external field), see Section A.2 in the appendix. We measure the speed of convergence of a Markov chain by the spectral gap, the difference between the largest eigenvalue, which is always one for stochastic matrices, and the second largest eigenvalue modulus λ_2 of the transition matrix (corresponding to $\boldsymbol{T}$). A spectral gap of value 1 means that the Markov chain converges in a single step, while a value of 0 means that it is not an ergodic chain. A value between 0 and 1 means that the Markov chain converges exponentially towards the stationary distribution with the error going down with $\mathcal{O}(|\lambda_2|^k)$, where k numbers the steps.[3] We compared the convergence speed of the three methods by constructing the full transition matrices and calculating the spectral gap directly for 3×3 Ising models with different coupling strengths J. We use both Ising models with periodic and non-periodic boundary conditions for the lattice structure. We tested the chessboard order and a linear order for sweeping sites, see Section A.2 in the appendix.

There was no difference for the convergence speed between chessboard and linear update order, therefore we only present the results for the chessboard order. In general, the convergence profile of Gibbs sampling is relatively simple. Without coupling ($J = 0$) the sites are independent and the Markov chain ends up in the correct distribution in a single step. The stronger the coupling gets, the slower Gibbs sampling converges. The graphs for the modified Metropolis operator contain a discontinuity at $J = 0$ by definition. With zero coupling the modified Metropolis operator behaves exactly like the Gibbs operator, converging in a single step corresponding to a spectral gap of 1. For small non-zero coupling strengths the modified Metropolis operator exhibits the same behavior as the classical Metropolis operator, both converge very slowly with a spectral gap close to 0.

The results for experiments in which the classical Metropolis algorithm does not converge are shown in the top plot in Figure 2: For a 3×3 Ising model with a periodic lattice structure the spectral gap for Metropolis sampling is 0, irrespective of coupling strength. That is, the Markov chain induced by the classical Metropolis algorithm with both chessboard and linear update order is reducible for this model. In contrast, the modified Metropolis algorithm converges in this setting as expected from Theorem 1. After the discontinuity at $J = 0$, the convergence speed improves quickly with rising coupling strength until it reaches a point where it converges slightly faster than Gibbs sampling, but approaches the same convergence speed at very strong couplings. The bottom plot in Figure 2 depicts the results for a case where both the classical and modified Metropolis converge very slowly for models with weak coupling, but converge better with rising coupling strength until they surpass Gibbs sampling. After they reach a point with maximal convergence speed, the convergence speed decreases again and all three sampling methods approach the same spectral gap for very high coupling strengths. Although the overall shapes of the curves are very similar for both versions of the Metropolis algorithm, the spectral gap of the modified Metropolis operator rises quicker, reaches its maximum

[3] Calculating the spectral gap limits our analysis to quite small models, because the size of the transition matrix depends on the number of states, which grows exponentially with the number of variables. However, the spectral gap quantifies the speed of convergence accurately.

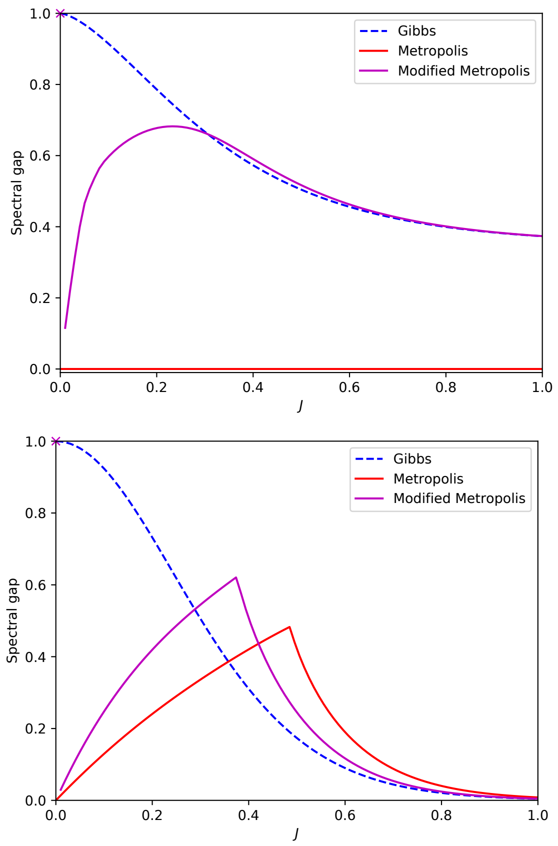

Figure 2: The spectral gap of the transition matrix for a 3 × 3 Ising model with periodic boundary (top) and non-periodic (bottom) conditions for the Gibbs (blue, dashed) as well as Metropolis (red) and modified Metropolis operator (magenta, dotted, discontinuity at $J = 0$ marked with ×) in dependence on the coupling strength J.

earlier and then also decreases earlier and faster. In summary, there are models with low J where Gibbs sampling converges the fastest, models with medium J where the modified Metropolis algorithm converges the fastest, and models with high J where the classical algorithm converges the fastest. Nonetheless, the modified and classical Metropolis operator show a very similar behavior overall.

5 CONCLUSIONS

There has been a gap between theory and practice in MCMC sampling using the Metropolis algorithm. The necessary mathematical conditions ensuring an ergodic Markov chain do not, in general, hold for multivariate binary distributions for the most popular variant of the Metropolis algorithm, which updates the states of the random variables in a fixed order, but this has not stopped the community from employing it. This paper shows that only a very small modification of the Metropolis algorithm is necessary to produce an ergodic Markov chain that guarantees convergence for all multivariate binary distributions.

Moreover, our results allow to identify scenarios in which the standard Metropolis algorithm for multivariate binary distributions using fixed-order updates may not converge and scenarios in which the algorithm is safe to use. Specifically, reducibility only occurs if there exist states that differ only in one variable and have the same probability under the stationary distribution. If no such states exist, our convergence proof can be directly transferred to the standard Metropolis algorithm – and this will almost surely be the case for typical models such as Boltzmann machines and spin glasses with (initial) parameters drawn from a continuous distribution. This insight closes the gap between practice and (Markov chain) theory for the Metropolis algorithm with fixed-order updates for multivariate binary distributions.

We measured the spectral gap for Ising models (with uniform coupling strength and no external field), where the standard Metropolis algorithm with fixed update order is not guaranteed to converge. In settings where the standard Metropolis MCMC converges, the modified Metropolis algorithm has a similar convergence speed profile as the original algorithm, but behaves slightly more like Gibbs sampling. When looking at Ising models where the standard Metropolis algorithms does not converge, the modified version does. Here Gibbs sampling gives better results for low coupling strengths, while the new Metropolis operator performs slightly better for medium coupling strengths.

In summary, we see no argument why the newly proposed transition operator should not be the default setting for Metropolis MCMC with fixed-order updates of binary variables.

Acknowledgments

CI acknowledges support by the Villum Foundation through the project Deep Learning and Remote Sensing for Unlocking Global Ecosystem Resource Dynamics (DeReEco).

A EXAMPLES WHERE THE METROPOLIS ALGORITHM DOES NOT CONVERGE

A.1 MINIMAL EXAMPLE

Let us consider binary $\Omega = \{0, 1\}$ and $n = 2$ variables and assume that the target distribution p is uniform, that is, $p((0, 0)) = p((0, 1)) = p((1, 0)) = p((1, 1))$. Assume we use the standard Metropolis algorithm with fixed-order updates and assume that we start in state $(0, 0)$. The algorithm will accept all proposed transitions for each variable and go from $(0, 0)$ to $(1, 1)$ and from there back to $(0, 0)$. The states $(1, 0)$ and $(0, 1)$ will never be reached. In contrast, the modified transition operator may reject a change, and each state can be reached.

In general, ergodicity is not an issue for variables with more than two states. For example, if we consider $\Omega = \{-1, 0, 1\}$, $n = 2$, and p uniform, then it is possible to go from any (x_1, x_2) to (y_1, y_2) in two steps via the state $(\Omega \setminus \{x_1, y_1\}, \Omega \setminus \{x_2, y_2\})$.

A.2 ISING MODELS

An Ising model describes a d-dimensional lattice with interactions between neighboring sites, where the coupling strength J controls how strong a tendency there is for neighbouring sites to be the same. We have $\Omega = \{-1, +1\}$ and denote the set of pairs of indexes of neighboring sites by Λ. The probability of the state of an Ising model is given by

$$p(\boldsymbol{x}) = \frac{e^{-E(\boldsymbol{x})}}{\sum_{\boldsymbol{x} \in \Omega^n} e^{-E(\boldsymbol{x})}} \text{ with } E(\boldsymbol{x}) = -J \sum_{(i,j) \in \Lambda} x_i x_j \ .$$

Counter-examples, where the Metropolis acceptance function with fixed-order updates leads to a reducible Markov chain, were found, for instance, by Friedberg and Cameron (1970), who simulated 4 × 4 two-dimensional Ising models. They discovered, that there can be “locked in configurations”, but did not make the connection to Markov chain theory and irreducibility. Two examples are shown in Figure A.3 and Figure A.4. Both examples consider *periodic* models, that is, we we assume periodic boundary conditions (i.e., sites on one edge of the lattice are connected to sites on the opposite edge).

B EXPERIMENTS WITH RANDOM SCANNING

As described in Section 2, the question whether random-order or fixed-order updates are preferable has been discussed in the literature and is not focus of our study. Random scanning can be used for all methods discussed in this paper, and here we include the results of rerunning all our experiments with random scans.

As can be seen from Figure B.5, the results generally support the popular belief that fixed-order scans lead to faster convergence: For almost all settings random-order scans show much worse performance than Gibbs sampling with fixed-order scans. The only exception is the setting investigating an Ising model with periodic boundary conditions, where classical Metropolis sampling with fixed-order scans does not converge, but performs exceptionally well with random-order scans for high J.

References

C. Andrieu. On random-and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.

P. Billingsley. *Measure and Probability.* Wiley, 2 edition, 1995.

K. Brügge, A. Fischer, and C. Igel. The flip-the-state transition operator for restricted Boltzmann machines. *Machine Learning*, 93(1):53–69, 2013.

S. G. Brush. History of the Lenz-Ising Model. *Reviews of Modern Physics*, 39:883–893, 1967.

G. W. Cunningham and P. H. E. Meijer. A comparison of two Monte Carlo methods for computations in statistical mechanics. *Journal of Computational Physics*, 20(1):50–63, 1974.

P. Diaconis. Some things we've learned (about Markov chain Monte Carlo). *Bernoulli*, 19(4):1294–1305, 2013.

A. Fischer and C. Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47: 25–39, 2014.

A. Fischer and C. Igel. A bound for the convergence rate of parallel tempering for sampling restricted Boltzmann machines. *Theoretical Computer Science*, 598:102–117, 2015.

R. Friedberg and J. E. Cameron. Test of the Monte Carlo method: fast simulation of a small Ising lattice. *The Journal of Chemical Physics*, 52(12):6049–6058, 1970.

A. Frigessi, P. Di Stefano, C.-R. Hwang, and S.-J. Sheu. Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–219, 1993.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation of North America, 1991.

H. Guo, K. Kara, and C. Zhang. Layerwise systematic scan: Deep Boltzmann machines and beyond. In

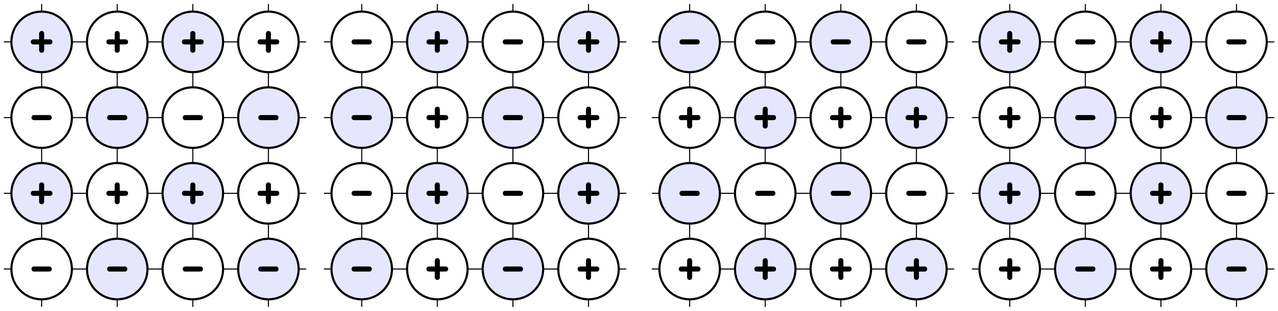

Figure A.3: Counter-example with chessboard update order. We assume a periodic Ising model with uniform coupling-strengths, where both of the dimensions of the Ising model have an even number of sites. If the site activations form a striped pattern and are updated in a chessboard pattern (here indicated by the coloring of the nodes, where first the nodes of one color are updated, then the nodes of the other color), they will never escape the striped pattern, switching deterministically from horizontal to vertical stripes after updating half the variables, switching to the complementary horizontal striped pattern after a full update step and switching back in another full update step.

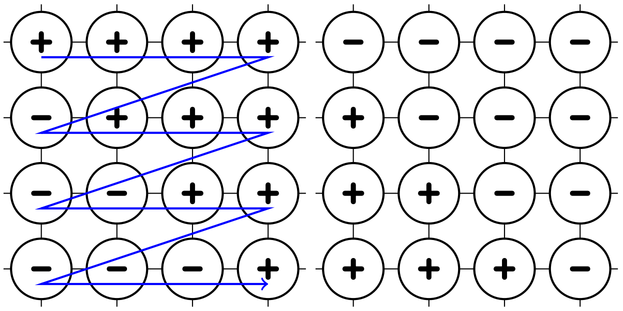

Figure A.4: Counter-example with left-to-right, top-to-bottom scanning order. For a periodic Ising model, where both of the dimensions have the same number of sites, if the site activations form a triangular pattern as shown in the left figure, they will flip between two possibles states if updated in the order indicated by the arrow.

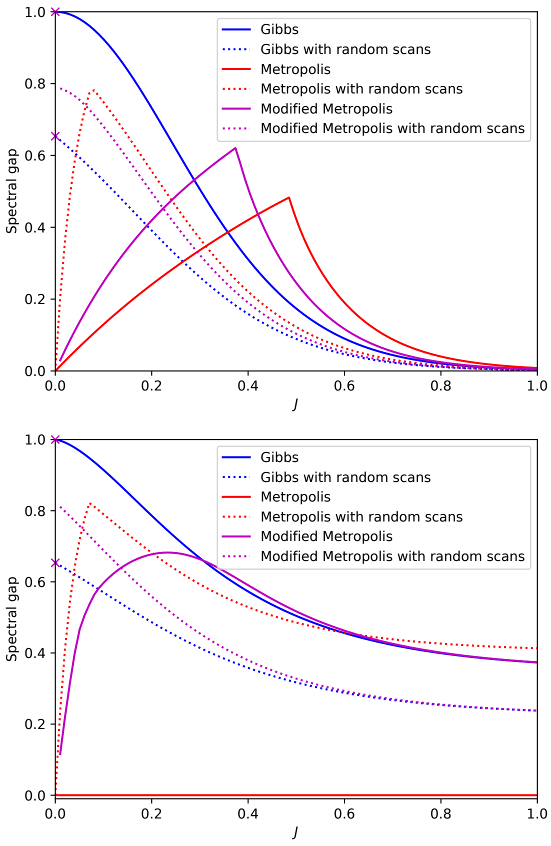

Figure B.5: Random scan results for non-periodic (left) and periodic (right) boundary conditions, compare to Figure 2.

International Conference on Artificial Intelligence and Statistics (AISTATS), pages 178–187, 2018.

B. D. He, C. M. De Sa, I. Mitliagkas, and C. Ré. Scan order in Gibbs sampling: Models in which it matters and bounds on how much. In *Advances in Neural Information Processing Systems*, pages 1–9, 2016.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

J. P. Hobert, A. Tan, and R. Liu. When is Eaton's Markov chain irreducible? *Bernoulli*, 13(3):641–652, 2007.

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

R. A. Levine. A note on Markov chain Monte Carlo sweep strategies. *Journal of Statistical Computation and Simulation*, 75(4):253–262, 2005.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of*

Chemical Physics, 21(6):1087–1092, 1953.

I. Mitliagkas and L. Mackey. Improving Gibbs sampler scan quality with DoGS. In *International Conference on Machine Learning (ICML)*, pages 2469–2477, 2017.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 1993.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

P. H. Peskun. Guidelines for choosing the transition matrix in Monte Carlo methods using Markov chains. *Journal of Computational Physics*, 40(2):327–344, 1981.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, 59(2):291–317, 1997.

R. R. Salakhutdinov. Learning in Markov random fields using tempered transitions. In *Advances in Neural Information Processing Systems*, pages 1598–1606, 2009.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*, pages 194–281. MIT Press, 1986.

R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters*, 58(2):86, 1987.

M. Wahlström. Euler digraphs. In J. Bang-Jensen and G. Gutin, editors, *Classes of Directed Graphs.* Springer, 2018.

U. Wolff. Collective monte carlo updating for spin systems. *Physical Review Letters*, 62(4):361, 1989.