Layerwise Systematic Scan: Deep Boltzmann Machines and Beyond

Heng Guo
University of Edinburgh

Kaan Kara
ETH Zurich

Ce Zhang
ETH Zurich

Abstract

For Markov chain Monte Carlo methods, one of the greatest discrepancies between theory and system is the *scan order* — while most theoretical development on the mixing time analysis deals with random updates, real-world systems are implemented with systematic scans. We bridge this gap for models that exhibit a bipartite structure, including, most notably, the Restricted/Deep Boltzmann Machine. The de facto implementation for these models scans variables in a layer-wise fashion. We show that the Gibbs sampler with a layer-wise alternating scan order has its relaxation time (in terms of epochs) no larger than that of a random-update Gibbs sampler (in terms of variable updates). We also construct examples to show that this bound is asymptotically tight. Through standard inequalities, our result also implies a comparison on the mixing times.

1 Introduction

Gibbs sampling, or the Markov chain Monte Carlo method in general, plays a central role in machine learning and has been widely implemented as the backbone algorithm for models such as Deep Boltzmann Machines (Salakhutdinov and Hinton, 2009), latent Dirichlet allocations (Blei et al., 2003), and factor graphs in general. Given a set of random variables and a target distribution π, the Gibbs sampler iteratively updates one variable at a time according to the distribution π conditioned on the values of all other variables. If the ergodicity condition is met, then the Gibbs sampler eventually converges to the target distribution.

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

There are two ways to choose which variable to update at the next iteration: (1) *Random Update*, where in each epoch (or round) one variable is picked uniformly at random with replacement; and (2) *Systematic Scan*, where in each epoch all variables are updated using some pre-determined order. Although most theoretical development on analyzing Gibbs sampling deals with random updates (Jerrum, 2003; Levin et al., 2009), systematic scans are prevalent in real-world implementations due to their hardware-friendly nature (cache locality for factor graphs, SIMD for Deep Boltzmann Machines, etc.). It is natural to wonder, *whether using systematic scan, rather than random updates, would delay the mixing time, the number of iterations the Gibbs sampler requires to reach the target distribution.*

The mixing time of these two update strategies can differ by some high polynomial factors in either directions (He et al., 2016; Roberts and Rosenthal, 2015). Even more pathological examples were constructed for non-Gibbs Markov chains such that systematic scan is not even ergodic whereas the random-update sampler is rapidly mixing (Dyer et al., 2008). Indeed, even for a system as simple as the Ising model, a comparison result remains elusive (Levin et al., 2009, Open problem 5, p. 300). As a consequence, theoretical results on rapid mixing, such as (Bubley and Dyer, 1997; Mossel and Sly, 2013), do not readily apply to the scan algorithms used in practice.

1.1 Main results.

In this paper, we bridge this gap between theory and system. We focus on *bipartite distributions*, in which variables can be divided into two partitions — conditioned on one of the partitions, variables from the other partition are mutually independent. This bipartite structure arises naturally in practice, including Restricted/Deep Boltzmann Machines. For a bipartite distribution, the de facto implementation is that in each epoch, we scan all variables from one of the partitions first, and then the other. We call this the *alternating-scan* sampler. Note that in order to define a valid Markov chain, we have to consider systematic

scans in epochs, in which all variables are updated once. Our main theorem is the following.

Theorem 1 (Main Theorem). *For any bipartite distribution π, if the random-update Gibbs sampler is ergodic, then so is the alternating-scan sampler. Moreover, the relaxation time of the alternating-scan sampler (in terms of epochs) is no larger than that of the random-update one (in terms of variable updates).*

The relaxation time (inverse spectral gap) measures the mixing time from a "warm" start. It is closely related to the (total variation distance) mixing time, and governs mixing times under other metrics as well (Levin et al., 2009). Through standard inequalities, Theorem 1 also implies a comparison result in terms of mixing times, Corollary 5. As we count epochs in Theorem 1, the alternating-scan sampler is implicitly slower by a factor of n, the number of variables. We also show that Theorem 1 is asymptotically tight via Example 6. Thus this implicit factor n slowdown cannot be improved in general.

More specifically, we summarize our contribution as follows.

1. In Section 4, we establish Theorem 1. By focusing on bipartite systems, we are able to obtain a much stronger result than recent studies in the more general setting (He et al., 2016). We note that standard Markov chain comparison results, such as (Diaconis and Saloff-Coste, 1993), do not seem to fit into our setting. Instead, we give a novel analysis via estimates of operator norms of certain carefully defined matrices. One key observation is to consider an artificial but equivalent variant of the alternating-scan sampler, where we insert an extra random update between updating variables from the two partitions. This does not change the algorithm since the extra random update is either redundant with the updates in the first partition or with those in the second.

2. In Section 5, we discuss bipartite distributions that arise naturally in machine learning. In particular, our result is a rigorous justification of the popular layer-wise scan sampler for Deep Boltzmann Machines (Salakhutdinov and Hinton, 2009). Our result also applies to other models such as Restricted Boltzmann Machines (Smolensky, 1986) and, more generally, any bipartite factor graph.

3. In Section 6, we conduct experiments to verify our theory and analyze the gap between our worst case theoretical bound and numerical evidences. We observe that in the rapidly mixing regime, the alternating-scan sampler is usually faster than the random-update one, whereas in the slow mixing regime, the alternating-scan sampler can be slower by a factor $O(n)$. We hope these observations shed some light on more fine-grained comparison bounds in the future.

2 Related Work

Probably the most relevant work is the recent analysis conducted by He et al. (2016) about the impact of the scan order on the mixing time of the Gibbs sampling. They (1) constructed a variety of models in which the scan order can change the mixing time significantly in several different ways and (2) proved comparison results on the mixing time between random updates and a variant of systematic scans where "lazy" moves are allowed. In this paper, we focus on a more specific case, i.e., bipartite systems, and so our bound is stronger — in fact, our bound can be exponentially stronger when the underlying chain is torpidly mixing. Moreover, our result does not modify the standard scan algorithm.

Another related work is the recent analysis by Tosh (2016) considering the mixing time of an alternating sampler for the Restricted Boltzmann Machine (RBM). Tosh showed that, under Dobrushin-like conditions (Dobrushin, 1970), i.e., when the weights in the RBM are sufficiently small, the alternating sampler mixes rapidly. For models other than RBM, mixing time results for systematic scans are relatively rare. Known examples are usually restricted to very specific models (Diaconis and Ram, 2000) or under conditions to ensure that the correlations are sufficiently weak (Dyer et al., 2006; Hayes, 2006; Dyer et al., 2008). Typical conditions of this sort are variants of the classical Dobrushin condition (Dobrushin, 1970). See also (Blanca et al., 2018) for very recent results on analyzing the alternating scan sampler (among others) on the 2D grid under conditions of the Dobrushin-type. In contrast, our work focuses on the relative performance between random updates and systematic scan, and does not rely on Dobrushin-like conditions. In particular, our results extend to the torpid mixing regime as well as the rapid mixing one.

Our primary focus is on discrete state spaces. The scan order question has also been asked and explored in general state spaces. Despite a long line of research (Hastings, 1970; Peskun, 1973; Caracciolo et al., 1990; Liu et al., 1995; Roberts and Sahu, 1997; Roberts and Rosenthal, 1997; Tierney, 1998; Maire et al., 2014; Roberts and Rosenthal, 2015; Andrieu, 2016), to the best of our knowledge, no decisive answer is known.

Another line of related research is about the scan order in *stochastic gradient descent* (Recht and Ré, 2012; Shamir, 2016; Gürbüzbalaban et al., 2017). Our set-

ting in this paper is very different and the techniques are different as well.

3 Preliminaries on Markov Chains

Let Ω be a discrete state space and P be a $|\Omega|$-by-$|\Omega|$ stochastic matrix describing a (discrete time) Markov chain on Ω. The matrix P is also called the transition matrix or the kernel of the chain. Thus, $P^t(\sigma_0, \cdot)$ is the distribution of the chain at time t starting from σ_0. Let $\pi(\cdot)$ be a stationary distribution of P. The Markov chain defined by P is *reversible* (with respect to $\pi(\cdot)$) if P satisfies the detailed balance condition:

$$\pi(\sigma)P(\sigma,\tau) = \pi(\tau)P(\tau,\sigma) \tag{1}$$

for any $\sigma, \tau \in \Omega$. We note that in general the systematic scan sampler is not reversible. The Markov chain is called *irreducible* if P connects the whole state space Ω, namely, for any $\sigma, \tau \in \Omega$, there exists t such that $P^t(\sigma,\tau) > 0$. It is called *aperiodic* if $\gcd\{t > 0 : P^t(\sigma,\sigma) > 0\} = 1$ for every $\sigma \in \Omega$. We call P *ergodic* if it is both irreducible and aperiodic. An ergodic Markov chain converges to its unique stationary distribution (Levin et al., 2009).

The *total variation* distance $\|\cdot\|_{TV}$ for two distributions μ and ν on Ω is defined as

$$\|\mu - \nu\|_{TV} = \max_{A\subset\Omega} |\mu(A) - \nu(A)| = \frac{1}{2}\sum_{\sigma\in\Omega} |\mu(\sigma) - \nu(\sigma)|.$$

The mixing time T_{mix} is defined as

$$T_{mix}(P) := \min\left\{t \geq 0 : \max_{\sigma\in\Omega} \left\|P^t(\sigma,\cdot) - \pi\right\|_{TV} \leq \frac{1}{2e}\right\},$$

where the choice of the constant $\frac{1}{2e}$ is merely for convenience and is not significant (Levin et al., 2009).

When P is ergodic and reversible, the eigenvalues $(\xi_i)_{i\in[|\Omega|]}$ of P satisfies $-1 < \xi_i \leq 1$, and additionally, $Pf = f$ if and only if f is constant (see (Levin et al., 2009, Lemma 12.1)). The *spectral gap* of P is defined by

$$\lambda(P) := 1 - \max\{|\xi| : \xi \text{ is an eigenvalue of } P \text{ and } \xi \neq 1\}. \tag{2}$$

The *relaxation time* for a reversible P is defined as

$$T_{rel}(P) := \lambda(P)^{-1}. \tag{3}$$

The relaxation time and the mixing time differ by at most a factor of $\log\left(\frac{2e}{\pi_{min}}\right)$ where $\pi_{min} = \min_{\sigma\in\Omega}\pi(\sigma)$, shown by the following theorem (see, for example, (Levin et al., 2009, Theorem 12.4 and 12.5)).

In fact, the relaxation time governs mixing properties with respect to metrics other than the total variation distance as well. See (Levin et al., 2009, Chapter 12) for more details.

Theorem 2. *Let P be the transition matrix of a reversible and ergodic Markov chain with the state space Ω and the stationary distribution π. Then*

$$T_{rel}(P) - 1 \leq T_{mix}(P) \leq T_{rel}(P)\log\left(\frac{2e}{\pi_{min}}\right),$$

where $\pi_{min} = \min_{\sigma\in\Omega}\pi(\sigma)$.

The factor $\log\pi_{min}^{-1}$ is usually bounded by a polynomial in the number of variables, in the context of Gibbs sampling which is our primary focus later. Theorem 2 is tight, and there is no good way of avoiding losing this $\log\pi_{min}^{-1}$ factor in general, with the spectral method.

Unfortunately, the systematic-scan sampler is not reversible, and therefore Theorem 2 does not apply. Instead, we use an extension developed by Fill (1991). For a non-reversible transition matrix P, let the *multiplicative reversiblization* be $R(P) := PP^*$, where P^* is the *adjoint* of P defined as

$$P^*(\sigma,\tau) = \frac{\pi(\tau)P(\tau,\sigma)}{\pi(\sigma)}. \tag{4}$$

Then $R(P)$ is reversible. Let the *relaxation time* for a (not necessarily reversible) P be

$$T_{rel}(P) := \frac{1}{1 - \sqrt{1 - \lambda(R(P))}}. \tag{5}$$

In particular, if P is reversible, then (5) recovers (3). In general, the multiplicative reversibilization mixes similarly to the original non-reversible chain. See (Fill, 1991) for more details.

The following theorem is a simple consequence of (Fill, 1991, Theorem 2.1).

Theorem 3. *Let P be the transition matrix of an ergodic Markov chain with the state space Ω and the stationary distribution π. Then*

$$T_{mix}(P) \leq \log\left(\frac{4e^2}{\pi_{min}}\right) T_{rel}(P),$$

where $\pi_{min} = \min_{\sigma\in\Omega}\pi(\sigma)$.

Note that our definition of relaxation times (5) for non-reversible Markov chains yields asymptotically the same upper bound in Theorem 2.

4 Alternating Scan

In this section we describe the random update and the alternating scan sampler, and compare these two.

Algorithm 1 Gibbs sampling with random updates

Input: Starting configuration $\sigma = \sigma_0$
  for $t = 1, \dots, T_{mix}$ **do**
    With probability $1/2$, do nothing.
    Otherwise, select a variable $x \in V$ uniformly at random.
    Set $\sigma \leftarrow \sigma^{x,s}$ with probability $\frac{\pi(\sigma^{x,s})}{\sum_{t \in S} \pi(\sigma^{x,t})}$.
  end for
  return σ

Let $V = \{x_1, \dots, x_n\}$ be a set of variables where each variable takes values from some finite set S. Let $\pi(\cdot)$ be a distribution defined on S^V.

Let $\sigma \in S^V$ be a configuration, namely $\sigma : V \to S$. Let $\sigma^{x,s}$ be the configuration that agrees with σ except at x, where $\sigma^{x,s}(x) = s$ for $s \in S$. In other words, for any $y \in V$,

$$\sigma^{x,s}(y) := \begin{cases} \sigma(y) & \text{if } y \neq x; \\ s & \text{if } y = x. \end{cases}$$

The lazy[1] Gibbs sampler is defined in Algorithm 1. Let $n = |V|$ be the total number of variables. The transition kernel P_{RU} (where RU stands for "random updates") of the sampler in Algorithm 1 is defined as:

$$P_{RU}(\sigma, \tau) = \begin{cases} \frac{1}{2n} \cdot \frac{\pi(\sigma^{x,s})}{\sum_{t \in S} \pi(\sigma^{x,t})} & \text{if } \tau \neq \sigma \text{ and there are } x \in V \\ & \quad \text{and } s \in S \text{ such that } \tau = \sigma^{x,s}; \\ 1/2 + \sum_{x \in V} \frac{1}{2n} \cdot \frac{\pi(\sigma^{x,\sigma(x)})}{\sum_{t \in S} \pi(\sigma^{x,t})} & \text{if } \tau = \sigma; \\ 0 & \text{otherwise,} \end{cases}$$

where σ, τ are two configurations. It is not hard to see, for example, by checking the detailed balance condition (1), that $\pi(\cdot)$ is the stationary distribution of P_{RU}. Note that this Markov chain is *lazy*, i.e., it remains at its current state with probability at least $1/2$. This self-loop probability is higher than $1/2$ because when we update a variable there is positive probability of no change. Lazy chains are often studied in the literature because of their technical conveniences. The self-loop eliminates potential periodicity, and all eigenvalues of a lazy chain are non-negative. In the context of Gibbs sampling, these are merely artifacts of the available techniques and considering the lazy version is not really necessary (Rudolf and Ullrich, 2013; Dyer et al., 2014). Our main result actually applies to both lazy and non-lazy versions. See the remarks after the proof of Theorem 1.

[1] We choose to present the lazy sampler due to its popularity in theoretical analysis. Our arguments later in fact also apply to non-lazy samplers as well. See the remarks after the proof of Theorem 1.

Algorithm 2 Alternating-scan sampler

Input: Starting configuration $\sigma = \sigma_0$
  for $t = 1, \dots, T_{mix}$ **do**
    for $i = 1, \dots, n_1$ **do**
      Set $\sigma \leftarrow \sigma^{x_i,s}$ with probability $\frac{\pi(\sigma^{x_i,s})}{\sum_{t \in S} \pi(\sigma^{x_i,t})}$.
    end for
    for $j = 1, \dots, n_2$ **do**
      Set $\sigma \leftarrow \sigma^{y_j,s}$ with probability $\frac{\pi(\sigma^{y_j,s})}{\sum_{t \in S} \pi(\sigma^{y_j,t})}$.
    end for
  end for
  return σ

Our main focus is bipartite distributions, defined next. These distributions arise naturally from bipartite factor graphs, including, most notably, Restricted Boltzmann Machines.

Definition 4. *The joint distribution $\pi(\cdot)$ of random variables $V = \{x_1, \dots, x_n\}$ is* bipartite, *if V can be partitioned into two sets V_1 and V_2 (namely $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$), such that conditioned on any assignment of variables in V_2, all variables in V_1 are mutually independent, and vice versa.*

In the following we consider a particular systematic scan sampler for bipartite distributions. For a configuration σ, let $\sigma_i := \sigma|_{V_i}$ be its projection on V_i where $i = 1, 2$. The alternating-scan sampler is given in Algorithm 2, where $n_1 = |V_1|$ and $n_2 = |V_2|$.

In other words, the alternating-scan sampler sequentially resamples all variables in V_1, and then resamples all variables in V_2. Note that since we are considering a bipartite distribution, in order to resample $x_i \in V_1$, we only need to condition on σ_2. In other words, for any $i \in [n_1]$, the distribution $\left(\frac{\pi(\sigma^{x_i,s})}{\sum_{t \in S} \pi(\sigma^{x_i,t})}\right)_{s \in S}$ that we draw from depends only on σ_2. Similarly, resampling $y_j \in V_2$ only depends on σ_1. We will denote the transition kernel of the alternating-scan sampler as P_{AS}, where AS stands for "alternating scan".

An unusual feature of systematic-scan samplers (including the alternating-scan sampler) is that they are not reversible. Namely the detailed balance condition (1) does not in general hold. This is because updating variables x and y in order is in general different from updating y and x in order. This imposes a technical difficulty as most of the theoretical tools for analyzing these chains are not suitable for irreversible chains, such as the Dirichlet form (Diaconis and Saloff-Coste, 1993) or conductance bounds (Jerrum and Sinclair, 1993; Sinclair, 1992).

On the other hand, the scan sampler is aperiodic. Any potential state σ of the chain must be in the state space Ω. Therefore $\pi(\sigma) > 0$ and the probability of staying

in σ is strictly positive. Moreover, if the Gibbs sampler is irreducible (namely the state space Ω is connected via single variable flips), then so is the scan sampler. This is because any single variable update can be simulated in the scan sampler, with small but strictly positive probability. Hence if the Gibbs sampler is ergodic, then so is the scan sampler.

We restate our main theorem here in formal terms.

Theorem 1. *For any bipartite distribution π, if P_{RU} is ergodic, then so is P_{AS}. Moreover,*

$$T_{rel}(P_{AS}) \le T_{rel}(P_{RU}).$$

Due to the space limit, we provide a proof sketch here. Complete details can be found in the supplementary material.

Proof sketch. The first statement is straightforward. For the second, let S_π be the projection matrix of the stationary distribution, namely

$$S_\pi(\sigma, \tau) = \pi(\tau).$$

If P is reversible, then we can rewrite the spectral gap in terms of an operator norm, namely,

$$\lambda(P) = 1 - \|P - S_\pi\|_\pi \,, \tag{6}$$

where $\|\cdot\|_\pi$ is the operator norm with respect to the distribution π. The transition matrix of updating a particular variable x is the following

$$T_x(\sigma, \tau) = \begin{cases} \frac{\pi(\sigma^{x,s})}{\sum_{s\in S}\pi(\sigma^{x,s})} & \text{if } \tau = \sigma^{x,s} \text{ for some } s \in S; \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, let I be the identity matrix that $I(\sigma, \tau) = \mathbb{1}(\sigma, \tau)$. Then we have that

$$P_{RU} = \frac{I}{2} + \frac{1}{2n}\sum_{x\in V} T_x; \quad P_{AS} = \prod_{i=1}^{n_1} T_{x_i} \prod_{j=1}^{n_2} T_{y_j}.$$

We consider an artificial but equivalent variant of P_{AS}, where after updating all variables in V_1, we do a random update according to P_{RU}, and then proceed to update all variables in V_2. This is equivalent to P_{AS} since the extra random update is either redundant with the updates in V_1 or with those in V_2. To put it formally,

$$P_{AS} = \prod_{i=1}^{n_1} T_{x_i} \cdot P_{RU} \cdot \prod_{j=1}^{n_2} T_{y_j}.$$

Denote $P_{AS1} := \prod_{i=1}^{n_1} T_{x_i}$ and $P_{AS2} := \prod_{j=1}^{n_2} T_{y_j}$. It is easy to verify that $\|P_{AS1}\|_\pi \le 1$ and $\|P_{AS2}\|_\pi \le 1$ as these two operators are self-adjoint and idempotent. Next, since S_π is an identity element, we verify that

$$\begin{aligned} P_{AS} - S_\pi &= P_{AS1}P_{RU}P_{AS2} - S_\pi \\ &= P_{AS1}\left(P_{RU} - S_\pi\right)P_{AS2}. \end{aligned} \tag{7}$$

Since the operator norm is sub-multiplicative, we have that

$$\begin{aligned} \|P_{AS} - S_\pi\|_\pi &= \|P_{AS1}\left(P_{RU} - S_\pi\right)P_{AS2}\|_\pi \\ &\le \|P_{AS1}\|_\pi \|P_{RU} - S_\pi\|_\pi \|P_{AS2}\|_\pi \\ &\le \|P_{RU} - S_\pi\|_\pi \,, \end{aligned} \tag{8}$$

where we use the fact that $\|P_{AS1}\|_\pi \le 1$ and $\|P_{AS2}\|_\pi \le 1$. However, (8) is not enough to conclude our proof, as (6) only applies to reversible Markov chains whereas the alternating-scan sampler is not necessarily reversible.

Instead, we establish a similar inequality for the multiplicative reversibilization $R(P_{AS}) = P_{AS}P_{AS}^*$ so as to conclude using (6). We verify that

$$P_{AS}^* = P_{AS2}P_{AS1}. \tag{9}$$

Using (9) and with a bit more technical details, we derive an analogue of (7) as follows:

$$\begin{aligned} &R(P_{AS}) - S_\pi = \\ &\quad P_{AS1}(P_{RU} - S_\pi)P_{AS2}P_{AS2}(P_{RU} - S_\pi)P_{AS1}. \end{aligned}$$

Then, completely analogously to (8), we have that

$$\|R(P_{AS}) - S_\pi\|_\pi \le \|P_{RU} - S_\pi\|_\pi^2 \,.$$

Now we can apply (6) and conclude using (5). □

Remark. *It is easy to check that the proof also works if we consider the non-lazy version of P_{RU}. To do so, we just replace $\frac{I}{2} + \frac{1}{2n}\sum_{x\in V} T_x$ with $\frac{1}{n}\sum_{x\in V} T_x$ and the rest of the proof goes through without changes.*

Remark. *Our argument can also handle the case of general state spaces, such as Gaussian variables, since the essential property we use is the commutativity of updating variables from the same partition. For general state spaces, in order to apply Theorem 1 on mixing times, we need to replace Theorem 2 and Theorem 3 with their continuous counterparts. See for example (Lawler and Sokal, 1988).*

Using Theorem 2 and Theorem 3, we translate Theorem 1 in terms of the mixing time.

Corollary 5. *For a Markov random field defined on a bipartite graph, let P_{RU} and P_{AS} be the transition kernels of the random-update Gibbs sampler and the alternating-scan sampler, respectively. Then,*

$$T_{mix}(P_{AS}) \le \log\left(\frac{4e^2}{\pi_{min}}\right)\left(T_{mix}(P_{RU}) + 1\right),$$

where $\pi_{min} = \min_{\sigma\in\Omega}\pi(\sigma)$.

Since n variables are updated in each epoch of P_{AS}, one might hope to strengthen Theorem 1 so that

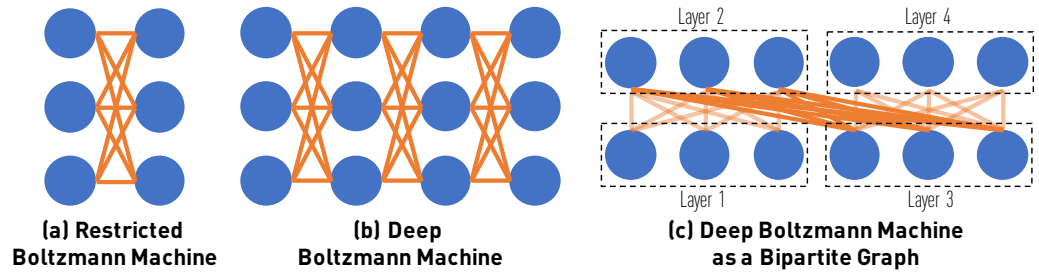

Figure 1: Restricted Boltzmann machines and deep Boltzmann machines as bipartite systems.

$nT_{rel}(P_{AS})$ is also no larger than $T_{rel}(P_{RU})$. Unfortunately, this is not the case and we give an example (similar to the "two islands" example due to He et al. (2016)) where $T_{mix}(P_{AS}) \asymp T_{mix}(P_{RU})$ and $T_{rel}(P_{AS}) \asymp T_{rel}(P_{RU})$. This example implies that Theorem 1 is asymptotically tight. However, it is still possible that Corollary 5 is loose by a factor of $\log \pi_{min}^{-1}$. This potential looseness is difficult to circumvent due to the spectral approach we took.

Example 6. *Let $G = (L \cup R, E)$ be a complete bipartite graph $K_{n,n}$ and we want to sample an uniform independent set in G. In other words, each vertex is a Boolean variable and a valid configuration is an independent set $I \subseteq L \cup R$. To be an independent set in $K_{n,n}$, I cannot intersect both L and R. Hence the state space is $\Omega = \{I \mid I \subseteq L \text{ or } I \subseteq R\}$ and the measure π is uniform on Ω. Under single-site updates, Ω is composed of two independent copies of the Boolean hypercube $\{0,1\}^n$ with the two origins identified. The random-update Gibbs sampler has mixing time $O(2^n)$ because the (maximum) hitting time of the Boolean hypercube is $O(2^n)$ and the mixing time is upper bounded by the hitting time multiplied by a constant (Levin et al., 2009, Eq. (10.24)). The relaxation time is also $O(2^n)$ by Theorem 2. In fact, it is not hard to see that both quantities are $\Theta(2^n)$.*

On the other hand, the alternating-scan sampler has mixing time $\Omega(2^n)$ and relaxation time $\Omega(2^n)$. For the mixing time, we partition the state space Ω into $\Omega_L = \{I \mid I \subset L\}$ and $\Omega_R = \{I \mid I \subset R \text{ and } I \neq \emptyset\}$. Consider the alternating scan projected down to Ω_L and Ω_R. If the current state is in Ω_L, then there is 2^{-n} probability to go to $\emptyset$ after updating all vertices in L, and then with probability $1 - 2^{-n}$ the state goes to Ω_R after updating all vertices in R. Similarly, going from Ω_R to Ω_L has also probability $O(2^{-n})$. Thus in each epoch of the alternating scan, the probability to go between Ω_L and Ω_R is $\Theta(2^n)$ and the mixing time is thus $\Theta(2^{-n})$. The relaxation time can be similarly bounded using a standard conductance argument (Sinclair, 1992).

In summary, for this bipartite distribution π, we have that $T_{rel}(P_{AS}) \asymp T_{rel}(P_{RU})$ and $T_{mix}(P_{AS}) \asymp T_{mix}(P_{RU})$. Therefore, Theorem 1 is asymptotically tight and Corollary 5 is tight up to the factor $\log \pi_{\min}^{-1}$.

We conjecture that the factor $\log \pi_{min}^{-1}$ should not be in Corollary 5. However, this factor is inherently there with the spectral approach. To get rid of it a new approach is required.

We note that in Example 6, alternating scan is not necessarily the best scan order. Indeed, as shown by He et al. (2016), if we scan vertices alternatingly from the left and right, rather than scanning variables layerwise, the mixing time is smaller by a factor of n. Thus, although Theorem 1 and Corollary 5 provide certain guarantees of the alternating-scan sampler, the layerwise alternating order is not necessarily the best one.

5 Bipartite Distributions in Machine Learning

The results we have developed so far can be applied to probabilistic graphic models with bipartite structures, most notably Restricted Boltzmann Machines (RBM) and Deep Boltzmann Machines (DBM). Although real-world systems for RBM and DBM inference rely on layerwise systematic scans, we are the first to provide a theoretical justification of such implementations.

5.1 Markov Random Fields

A Markov random field (MRF) with binary factors $\langle G, S, \pi \rangle$ is defined on a graph $G = (V, E)$, where each edge describes a "factor" f_e and each vertex is a variable drawing from S, a set of possible values. Each factor is a function $S^2 \to \mathbb{R}$. A configuration $\sigma \in S^V$ is a mapping from V to S. In addition, each vertex is equipped with a factor $g_v : S \to \mathbb{R}$. Let $\Omega \subseteq S^V$ be the state space, which is usually defined by a set of hard constraints. When there is no hard constraint, the state space Ω is simply S^V. The Hamiltonian of

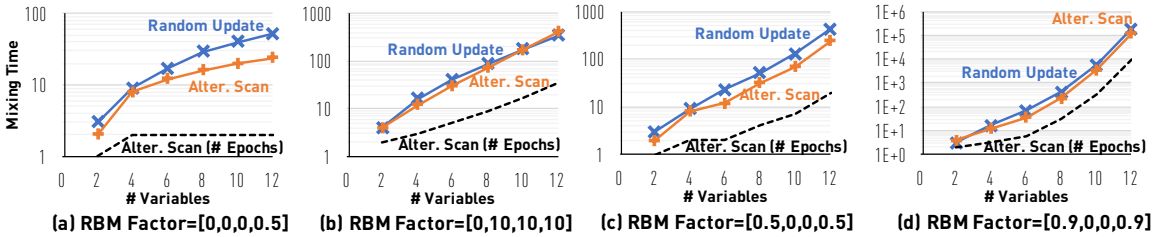

Figure 2: Mixing time of Gibbs samplers on Restricted Boltzmann Machines. See Section 6.

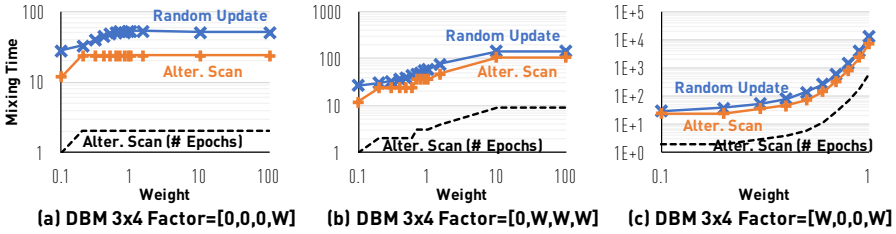

Figure 3: Mixing time of Gibbs samplers on Deep Boltzmann Machines. See Section 6.

$\sigma \in \Omega$ is defined as

$$H(\sigma) = \sum_{e=(u,v)\in E} f_e(\sigma(u), \sigma(v)) + \sum_{v\in V} g_v(\sigma(v)).$$

The Gibbs distribution $\pi(\cdot)$ is defined as $\pi(\sigma) \propto \mathbb{1}(\sigma \in \Omega)\exp(H(\sigma))$. These models are popularly used in applications such as image processing (Li, 2009) and natural language processing (Lafferty et al., 2001).

It is easy to check that, when the underlying graph G is bipartite, the Gibbs distribution is bipartite in the sense of Definition 4. Thus Theorem 1 and Corollary 5 apply to this setting.

5.2 Restricted/Deep Boltzmann Machines

Restricted Boltzmann Machines (RBM) was introduced by Smolensky (1986). It is a special case of the general MRF in which all variables are Boolean (i.e., $S = \{0, 1\}$) and are partitioned into two disjoint sets, V_1 and V_2. There is a factor between each variable in V_1 and V_2, and the Hamiltonian is

$$H(\sigma) = \sum_{u\in V_1, v\in V_2} W_{uv}\sigma(u)\sigma(v) + \sum_{v\in V} W_v\sigma(v).$$

where W_{uv} and W_v are real-valued weights. Figure 1(a) illustrates the structure of RBMs. We use $[f_{00}, f_{01}, f_{10}, f_{11}]$ to describe a general binary factor defined on Boolean variables. Thus, $[0, 0, 0, W]$ denotes a standard RBM factor with weight W, and $[W, 0, 0, W]$ denotes an Ising model with weight W (after some renormalization).

Markov chain Monte Carlo is a common approach to perform inference for RBMs, which involves sampling a configuration from the Gibbs distribution π. The de facto algorithm for this task is Gibbs sampling, in which the conditional probability of each step can be calculated from only the Hamiltonian. In this context, the alternating-scan algorithm we study corresponds to a *layerwise scan* — first update all variables in V_1 and then all variables in V_2. This scan order allows one to use efficient linear algebra primitives such as dense matrix multiplication implemented with GPUs or SIMD instructions on modern CPUs.

Deep Boltzmann Machines (DBM) (Salakhutdinov and Hinton, 2009) is a Deep Learning model that extends RBM to multiple layers as illustrated in Figure 1(b). This layer structure is indeed bipartite, shown in Figure 1(c). The scan order induced is thus to update odd layers first and even ones after. Like most deep learning models, the scan (evaluation) order of variables has significant impact on the speed and performance of the system. The layerwise implementation is particularly advantageous thanks to dense linear algebra primitives.

Given an RBM or DBM with n variables, it is easy to see that $\log \pi_{\min}^{-1}$ is $O(n)$. Thus, Corollary 5 implies that, comparing to the random-update algorithm, the layerwise systematic scan algorithm incurs at most a $O(n^2)$ slowdown in the convergence rate. Our result improves exponentially (in the worst case) upon the previous result by He et al. (2016).

6 Experiments

Empirically evaluating the mixing time of Markov chains is notoriously difficult. In general, it is hard under certain complexity assumptions (Bhatnagar et al., 2011) and lower bounds have been established for more concrete settings by Hsu et al. (2015) (see also (Hsu et al., 2015) for a comprehensive survey on this topic). We evaluate the mixing time in either exact and straightforward or approximate but tractable ways, including (1) calculating directly using the transition matrix for small graphs, (2) taking advantage of symmetries in the state space for medium-sized graphs, and (3) using the coupling time (defined later) as a proxy of the mixing time for large graphs.

Mixing Time on Small Graphs. We evaluate the mixing time in a brute force way, namely, we multiply the transition matrix until the total variation distance to the stationary distribution is below the threshold. Since the state space is exponentially large, such a method is only feasible in small graphs.

Figure 2 and Figure 3 contains the comparison of the mixing time for small graphs (RBMs of up to 12 variables and DBMs with 4 layers and 3 variable per layer). We vary (1) number of variables, (2) factor functions (shown as the entries of truth table in the caption), or (3) the weight of factors, in different figures and report the mixing times of random updates and layerwise scan. All solid lines count mixing time in # variable updates and the dotted line in # epochs.

We see that, empirically, alternating scan has comparable, sometimes better, mixing time than random updates, even when counting in the number of variable updates. On one hand, it confirms our result that the mixing time of alternating scan and random updates are similar. On the other, it shows that our result, although asymptotically tight for the worst case, is not "instance optimal". This observation indicates promising future direction for beyond-worst case analysis.

Medium-sized Graphs. We now turn to Example 6, which has also been studied by He et al. (2016) and is asymptotically the worst case of Theorem 1. Due to certain symmetries, we have a much more succinct representation of the state space, and manage to calculate the mixing and relaxation times for mildly larger graphs (up to 50 variables). As illustrated in Figure 4, the alternating-scan sampler is slower than, but still comparable to the random-update sampler. This is consistent with the discussion in Example 6.

Coupling Time on Large Graphs. Lastly, we use the coupling time as a proxy of the mixing time and estimate it on large graphs with 10^4 variables and 5×10^4 randomly chosen factors.

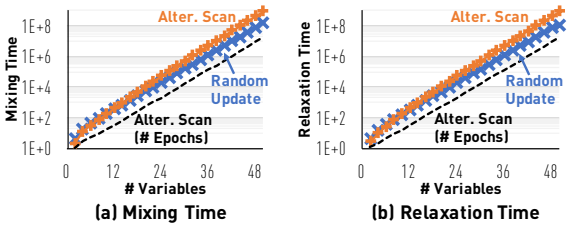

Figure 4: Mixing time comparison on medium-sized graphs.

We use the grand coupling (Levin et al., 2009, Chapter 5). Let $T_{\sigma,\tau}$ be the first time two copies of the same Markov chain meet, with initial states σ and τ, under certain coupling. Then the coupling time is $\max_{(\sigma,\tau)\in\Omega^2} T_{\sigma,\tau}$. All of the models we tested are monotone (Peres and Winkler, 2013), in which the coupling time under the grand coupling can be easily evaluated by simulating from the top and bottom states. The coupling time is closely related to the mixing time (Levin et al., 2009, Chapter 5). In fact, designing a good coupling is an important technique to proving rapid mixing (Bubley and Dyer, 1997).

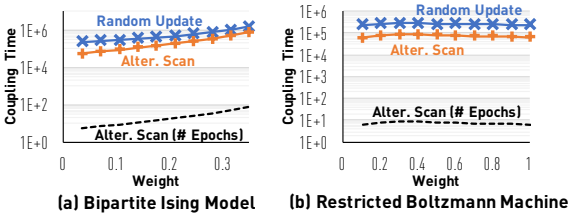

Figure 5: Coupling time comparison on large and random graphs.

In these experiments, we choose our parameters to stay within the rapidly mixing regime (Mossel and Sly, 2013). As we can see in Figure 5, alternating scan is faster than random updates (in terms of variable updates). Indeed, numerical evidence suggests that the speedup factor is close to 2.

Acknowledgement

CZ and the DS3Lab gratefully acknowledge the support from the Swiss National Science Foundation NRP 75 407540_167266, IBM Zurich, Mercedes-Benz Research & Development North America, Oracle Labs, Swisscom, Zurich Insurance, Chinese Scholarship Council, and the Department of Computer Science at ETH Zurich, the GPU donation from NVIDIA Corporation, the cloud computation resources from Microsoft Azure for Research award program.

References

Christophe Andrieu. On random- and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016. 2

Nayantara Bhatnagar, Andrej Bogdanov, and Elchanan Mossel. The computational complexity of estimating MCMC convergence time. In *RANDOM*, pages 424–435, 2011. 8

Antonio Blanca, Pietro Caputo, Alistair Sinclair, and Eric Vigoda. Spatial mixing and non-local Markov chains. *SODA*, 2018. *To appear*. 2

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, 2003. 1

Russ Bubley and Martin E. Dyer. Path coupling: A technique for proving rapid mixing in Markov chains. In *FOCS*, pages 223–231, 1997. 1, 8

Sergio Caracciolo, Andrea Pelissetto, and Alan D. Sokal. Nonlocal Monte Carlo algorithm for self-avoiding walks with fixed endpoints. *J. Stat. Phys.*, 60(1):1–53, Jul 1990. 2

Persi Diaconis and Arun Ram. Analysis of systematic scan Metropolis algorithms using Iwahori-Hecke algebra techniques. *Michigan Math. J.*, 48(1):157–190, 2000. 2

Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730, 08 1993. 2, 4

Roland L. Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.*, (3):458–486, 1970. 2

Martin Dyer, Catherine Greenhill, and Mario Ullrich. Structure and eigenvalues of heat-bath Markov chains. *Linear Algebra Appl.*, 454:57 – 71, 2014. 4

Martin E. Dyer, Leslie Ann Goldberg, and Mark Jerrum. Systematic scan for sampling colourings. *Ann. Appl. Probab.*, 16(1):185–230, 2006. 2

Martin E. Dyer, Leslie Ann Goldberg, and Mark Jerrum. Dobrushin conditions and systematic scan. *Combin. Probab. Comput.*, 17(6):761–779, 2008. 1, 2

James A. Fill. Eigenvalue bounds on convergence to stationary for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.*, 1(1):62–87, 1991. 3, 11

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Convergence rate of incremental aggregated gradient algorithms. *SIAM J. Optimiz.*, 2017. *To appear*. 2

Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, pages 97–109, 1970. 2

Thomas P. Hayes. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *FOCS*, pages 39–46, 2006. 2

Bryan D. He, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Scan order in Gibbs sampling: Models in which it matters and bounds on how much. In *NIPS*, pages 1–9, 2016. 1, 2, 6, 7, 8

Daniel J. Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. In *NIPS*, pages 1459–1467, 2015. 8

Mark Jerrum. *Counting, Sampling and Integrating: Algorithms and Complexity*. Lectures in Mathematics, ETH Zürich. Birkhäuser, 2003. 1

Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22(5):1087–1116, 1993. ISSN 0097-5397. 4

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001. 7

Gregory F. Lawler and Alan D. Sokal. Bounds on the l^2 spectrum for Markov chains and markov processes: A generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, 309(2):557–580, 1988. 5

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. 1, 2, 3, 6, 8

Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. 2009. 7

Jun S. Liu, Wing H. Wong, and Augustine Kong. Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Royal Stat. Soc. B*, 57(1):157–169, 1995. 2

Florian Maire, Randal Douc, and Jimmy Olsson. Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Stat.*, 42(4):1483–1510, 08 2014. 2

Elchanan Mossel and Allan Sly. Exact thresholds for Ising-Gibbs samplers on general graphs. *Ann. Probab.*, 41(1):294–328, 2013. 1, 8

Yuval Peres and Peter Winkler. Can extra updates delay mixing? *Comm. Math. Phys.*, 323(3):1007–1016, 2013. 8

Peter H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973. 2

Benjamin Recht and Christopher Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In *COLT*, pages 11.1–11.24, 2012. 2

Gareth O. Roberts and Jeffrey S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.*, 2:13–25, 1997. 2

Gareth O. Roberts and Jeffrey S. Rosenthal. Surprising convergence properties of some simple Gibbs samplers under various scans. *Int. J. Stat. Probab.*, 5(1), 2015. 1, 2

Gareth O. Roberts and Sujit K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *J. Royal Stat. Soc. B*, 59(2):291–317, 1997. 2

Daniel Rudolf and Mario Ullrich. Positivity of hit-and-run and related algorithms. *Electron. Commun. Probab.*, 18:49.1–49.8, 2013. 4

Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *AISTATS*, pages 448–455, 2009. 1, 2, 7

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *NIPS*, pages 46–54, 2016. 2

Alistair Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Comb. Probab. Comp.*, 1:351–370, 1992. 4, 6

Paul Smolensky. Parallel distributed processing: Explorations in the microstructure of cognition. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, pages 194–281, 1986. 2, 7

Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1): 1–9, 1998. 2

Christopher Tosh. Mixing rates for the alternating Gibbs sampler over restricted Boltzmann machines and friends. In *ICML*, pages 840–849, 2016. 2

Mario Ullrich. Rapid mixing of Swendsen-Wang dynamics in two dimensions. *Dissertationes Mathematicae*, 502, 2014. 12