HOGWILD!-Gibbs Can Be PanAccurate

Constantinos Daskalakis*
EECS & CSAIL, MIT
costis@csail.mit.edu

Nishanth Dikkala†
EECS & CSAIL, MIT
nishanthd@csail.mit.edu

Siddhartha Jayanti ‡
EECS & CSAIL, MIT
jayanti@mit.edu

November 27, 2018

Abstract

Asynchronous Gibbs sampling has been recently shown to be fast-mixing and an accurate method for estimating probabilities of events on a small number of variables of a graphical model satisfying Dobrushin's condition [DSOR16]. We investigate whether it can be used to accurately estimate expectations of functions of *all the variables* of the model. Under the same condition, we show that the synchronous (sequential) and asynchronous Gibbs samplers can be coupled so that the expected Hamming distance between their (multivariate) samples remains bounded by $O(\tau \log n)$, where n is the number of variables in the graphical model, and τ is a measure of the asynchronicity. A similar bound holds for any constant power of the Hamming distance. Hence, the expectation of any function that is Lipschitz with respect to a power of the Hamming distance, can be estimated with a bias that grows logarithmically in n. Going beyond Lipschitz functions, we consider the bias arising from asynchronicity in estimating the expectation of polynomial functions of all variables in the model. Using recent concentration of measure results [DDK17, GLP17, GSS18], we show that the bias introduced by the asynchronicity is of smaller order than the standard deviation of the function value already present in the true model. We perform experiments on a multi-processor machine to empirically illustrate our theoretical findings.

1 Introduction

The increasingly ambitious applications of data analysis, and the corresponding growth in the size of the data that needs to processed has brought important scalability challenges to machine learning algorithms. Fundamental methods such as Gradient Descent and Gibbs sampling, which were designed with a sequential computational model in mind, are to be applied on datasets of increasingly larger size. As such, there has recently been increased interest towards developing techniques for parallelizing these methods. However, these algorithms are inherently sequential and are difficult to parallelize.

HOGWILD!-SGD, proposed by Niu et al. [NRRW11], is a lock-free asynchronous execution of stochastic gradient descent that has been shown to converge under the right sparsity conditions. Several variants of this method, and extensions of the asynchronous execution approach have been recently proposed, and have found successful applications in a broad range of applications ranging from PageRank approximation, to deep learning and recommender systems [YHSD12, NO14, MBDC15, MPP+15, LWR+15, LWR+15, DSZOR15].

Similar to HOGWILD!-SGD, lock-free asynchronous execution of Gibbs sampling, called HOGWILD!-Gibbs, was proposed by Smola and Narayanamurthy [SN10], and empirically shown to work well on several models [ZR14]. Johnson et al. [JSW13] provide sufficient conditions under which they show theoretically that HOGWILD!-Gibbs produces samples with the correct mean in Gaussian models, while Terenin et al. [TSD15] propose a modification to the algorithm that is shown to converge under some strong assumptions

*Supported by NSF CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.
†Supported by NSF CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.
‡Supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

on asynchronous computation.

Input: Set of variables V, Configuration $x_0 \in S^{|V|}$, Distribution π
initialization;
for $t = 1$ **to** T **do**
  Sample i uniformly from $\{1, 2, \ldots, n\}$;
  Sample $X_i \sim \Pr_\pi [.|X_{-i} = x_{-i}]$ and set $x_{i,t} = X_i$;
  For all $j \neq i$, set $x_{j,t} = x_{j,t-1}$;
end

Algorithm 1: Gibbs Sampling

In a more recent paper, De Sa et al. [DSOR16] propose the study of HOGWILD!-Gibbs under a stochastic model of asynchronicity in graphical models with discrete variables. Whenever the graphical model satisfies Dobrushin's condition, they show that the mixing time of the asynchronous Gibbs sampler is similar to that of the sequential (synchronous) one. Moreover, they establish that the asynchronous Gibbs sampler accurately estimates probabilities of events on a sublinear number of variables, in particular events on up to $O(\varepsilon n/\log n)$ variables can be estimated within variational distance ε, where n is the total number of variables in the graphical model (Lemma 2, [DSOR16]).

Our Results. Our goal in this paper is to push the theoretical understanding of HOGWILD!-Gibbs to estimate functions of *all the variables in a graphical model*. In particular, we are interested in whether HOGWILD!-Gibbs can be used to accurately estimate the expectations of such functions. Results from [DSOR16] imply that an accurate estimation is possible whenever the function under consideration is Lipschitz with a good Lipschitz constant with respect to the Hamming metric. Under the same Dobrushin condition used in [DSOR16] (see Definition 3), and under a stochastic model of asynchronicity with weaker assumptions (see Section 2.1), we show that you can do better than the bounds implied by [DSOR16] even for functions with bad Lipschitz constants. For instance, consider quadratic functions on an Ising model, which is a binary graphical model, and serves as a canonical example of Markov random fields [LPW09, MS10, Fel04, DMR11, GG86, Ell93]. Under appropriate normalization, these functions take values in the range $[-n^2, n^2]$ and have a Lipschitz constant of n. Given this, the results of [DSOR16] would imply we can estimate quadratic functions on the Ising model within an error of $O(n)$. We improve this error to be of $O(\sqrt{n})$. In particular, we show the following in our paper:

- Starting at the same initial configuration, the executions of the sequential and the asynchronous Gibbs samplers can be coupled so that the expected Hamming distance between the multivariate samples that the two samplers maintain is bounded by $O(\tau \log n)$, where n is the number of variables in the graphical model, and τ is a measure of the average contention in the asynchronicity model of Section 2.1. See Lemma 2. More generally, the expectation of the d-th power of the Hamming distance is bounded by $C(d, \tau) \log^d n$, for some function $C(d, \tau)$. See Lemma 3.

- It follows from Lemmas 2 and 3 that, if a function f of the variables of a graphical model is K-Lipschitz with respect to the d-th power of the Hamming distance, then the bias in the expectation of f introduced by HOGWILD!-Gibbs under the asynchronicity model of Section 2.1 is bounded by $K \cdot C(d, \tau) \log^d n$. See Corollary 1.

- Next, we improve the bounds of Corollary 1 for functions that are degree-d polynomials of the variables of the graphical model. Low degree polynomials on graphical models are a natural class of functions which are of interest in many statistical tasks performed on graphical models (see, for instance, [DDK18]). For simplicity we show these improvements for the Ising model, but our results are extendible to general graphical models. We show, in Theorem 6, that the bias introduced by HOGWILD!-Gibbs in the expectation of a degree-d polynomial of the Ising model is bounded by $O((n \log n)^{(d-1)/2})$. This bound improves upon the bound computed by Corollary 1 by a factor of about $(n/\log n)^{(d-1)/2}$, as the Lipschitz constant with respect to the Hamming distance of a degree-d polynomial of the Ising model can be up to $O(n^{d-1})$. Importantly, the bias of $O((n \log n)^{(d-1)/2})$ that we show is introduced by the asynchronicity is of a lower order of magnitude than the standard deviation of degree-d polynomials of the Ising model, which is $O((n)^{d/2})$—see Theorem 3, and which is already experienced by the

sequential sampler. Moreover, in Theorem 7, we also show that the asynchronous Gibbs sampler is not adding a higher order variance to its sample. Thus, our results suggest that running Gibbs sampling asynchronously leads to a valid bias-variance tradeoff.

Our bounds for the expected Hamming distance between the sequential and the asynchronous Gibbs samplers follow from coupling arguments, while our improvements for polynomial functions of Ising models follow from a combination of our Hamming bounds and recent concentration of measure results for polynomial functions of the Ising model [DDK17, GLP17, GSS18].

- In Section 5, we illustrate our theoretical findings by performing experiments on a multi-core machine. We experiment with graphical models over two kinds of graphs. The first is the $\sqrt{n} \times \sqrt{n}$ grid graph (which we represent as a torus for degree regularity) where each node has 4 neighbors, and the second is the clique over n nodes.
  We first study how valid the assumptions of the asynchronicity model are. The main assumption in the model was that the average contention parameter τ doesn't grow as the number of nodes in the graph grows. It is a constant which depends on the hardware being used and we observe that this is indeed the case in practice. The expected contention grows linearly with the number of processors on the machine but remains constant with respect to n (see Figures 1 and 2).
  Next, we look at quadratic polynomials over graphical models associated with both the grid and clique graphs. We estimate their expected values under the sequential Gibbs sampler and HOGWILD!-Gibbs and measure the bias (absolute difference) between the two. Our theory predicts that this should scale at $\sqrt{n}$ and we observe that this is indeed the case (Figure 3). Our experiments are described in greater detail in Section 5.

2 The Model and Preliminaries

In this paper, we consider the Gibbs sampling algorithm as applied to discrete graphical models. The models will be defined on a graph $G = (V, E)$ with $|V| = n$ nodes and will represent a probability distribution π. We use S to denote the range of values each node in V can take. For any configuration $X \in S^{|V|}$, $\pi_i(.|X^{-i})$ will denote the conditional distribution of variable i given all other variables of state X.

In Section 4, we will look at Ising models, a particular class of discrete binary graphical models with pairwise local correlations. We consider the Ising model on a graph $G = (V, E)$ with n nodes. This is a distribution over $\Omega = \{\pm 1\}^n$, with a parameter vector $\vec{\theta} \in \mathbb{R}^{|V|+|E|}$. $\vec{\theta}$ has a parameter corresponding to each edge $e \in E$ and each node $v \in V$. The probability mass function assigned to a string x is

$$P(x) = \exp\left(\sum_{v \in V} \theta_v x_v + \sum_{e=(u,v) \in E} \theta_e x_u x_v - \Phi(\vec{\theta})\right),$$

where $\Phi(\vec{\theta})$ is the log-partition function for the distribution. We say an Ising model has *no external field* if $\theta_v = 0$ for all $v \in V$. For ease of exposition we will focus on the case with no external field in this paper. However, the results extend to Ising models with external fields when the functions under consideration (in Section 4) are appropriately chosen to be *centered*. See [DDK17].

Throughout the paper we will focus on bounded functions defined on the discrete space $S^{|V|}$. For a function f, we use $\|f\|_\infty$ to denote the maximum absolute value of the function over its domain. We will use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. In Section 4, we will study polynomial functions over the Ising model. Since $x_i^2 = 1$ always in an Ising model, any polynomial function of degree d can be represented as a multilinear function of degree d and we will refer to them interchangeably in the context of Ising models.

Definition 1 (Polynomial/Multilinear Functions of the Ising Model)**.** *A* degree-d polynomial *defined on n variables $x_1, \ldots, x_n$ is a function of the following form*

$$\sum_{S \subseteq [n]: |S| \leq d} a_S \prod_{i \in S} x_i,$$

where $a : 2^{[n]} \to \mathbb{R}$ is a coefficient vector.

When the degree $d = 1$, we will refer to the function as a linear function, and when the degree $d = 2$ we will call it a bilinear function. Note that since $X_u \in \{\pm 1\}$, any polynomial function of an Ising model is a multilinear function. We will use a to denote the coefficient vector of such a multilinear function and $\|a\|_\infty$ to denote the maximum element of a in absolute value. Note that we will use permutations of the subscripts to refer to the same coefficient, i.e., a_{ijk} is the same as a_{jik}. Also we will use the term d-linear function to refer to a multilinear function of degree d.
At times, for simplicity we will instead consider degree d polynomials of the form $f_a(x) = \sum_{i_1 i_2, \ldots, i_d} a_{i_1 i_2 \ldots i_d} x_{i_1} x_{i_2} \ldots x_{i_d}$. This form involves multiplicity in the terms, i.e. each monomial might appear multiple times. We can map from degree d polynomials without multiplicity of terms to functions of the above form in a straightforward manner by dividing each coefficient $a_{i_1 i_2 \ldots i_d}$ by an appropriate constant which captures how many times the term appears in the above notation. This constant lies between $d!$ and d^d.

We now give a formal definition of Dobrushin's uniqueness condition, also known as the high-temperature regime. First we define the influence of a node j on a node i.

Definition 2 (Influence in Graphical Models)**.** *Let π be a probability distribution over some set of variables V. Let B_j denote the set of state pairs (X, Y) which differ only in their value at variable j. Then the influence of node j on node i is defined as*

$$I(j, i) = \max_{(X,Y) \in B_j} \left\| \pi_i(.|X^{-i}) - \pi_i(.|Y^{-i}) \right\|_{TV}$$

Now, we are ready to state Dobrushin's condition.

Definition 3 (Dobrushin's Uniqueness Condition)**.** *Consider a distribution π defined on a set of variables V. Let*

$$\alpha = \max_{i \in V} \sum_{j \in V} I(j, i)$$

π is said to satisfy Dobrushin's uniqueness condition is $\alpha < 1$.

We have the following result from [DSOR16] about mixing time of Gibbs sampler for a model satisfying Dobrushin's condition.

Theorem 1 (Mixing Time of Sequential Gibbs Sampling)**.** *Assume that we run Gibbs sampling on a distribution that satisfies Dobrushin's condition, $\alpha < 1$. Then the mixing time of sequential-Gibbs is bounded by*

$$t_{mix-hog(\varepsilon)} \leq \frac{n}{1 - \alpha} \log \left(\frac{n}{\varepsilon} \right).$$

Definition 4. *For any discrete state space $S^{|V|}$ over the set of variables V, The* Hamming distance *between $x, y \in S^{|V|}$ is defined as $d_H(x, y) = \sum_{i \in V} \mathbb{1}_{\{x_i \neq y_i\}}$.*

Definition 5 (The greedy coupling between two Gibbs Sampling chains)**.** *Consider two instances of Gibbs sampling associated with the same discrete graphical model π over the state space $S^{|V|}$: $X_0, X_1, \ldots$ and $Y_0, Y_1, \ldots$. The following coupling procedure is known as the* greedy coupling*. Start chain 1 at X_0 and chain 2 at Y_0 and in each time step t, choose a node $v \in V$ uniformly at random to update in both the chains. Without loss of generality assume that $S = \{1, 2, \ldots, k\}$. Let $p(i_1)$ denote the probability that the first chain sets $X_{t,v} = i_1$ and let $q(i_2)$ be the probability that the second chain sets $Y_{t,v} = i_2$. Plot the points $\sum_{j=1}^{i} p(j) = P(i)$, and $\sum_{j=1}^{i} q(j) = Q(i)$ for all $i \in [k]$ on the interval from $[0, 1]$. Also pick $P(0) = Q(0) = 0$ and $P(k + 1) = Q(k + 1) = 1$. Couple the updates according to the following rule:*

Draw a number x uniformly at random from $[0, 1]$. Suppose $x \in [P(i_1), P(i_1 + 1)]$ and $x \in [Q(i_2), Q(i_2 + 1)]$. Choose $X_{t,v} = i_1$ and $Y_{t,v} = i_2$.

We state an important property of this coupling which holds under Dobrushin's condition, in the following Lemma.

Lemma 1. *The greedy coupling (Definition 5) satisfies the following property. Let $X_0, Y_0 \in S^{|V|}$ and consider two executions of Gibbs sampling associated with distribution π and starting at X_0 and Y_0 respectively. Suppose the executions were coupled using the greedy coupling. Suppose in the step $t = 1$, node i is chosen to be updated in both the models. Then,*

$$\Pr\left[X_{i,1} \neq Y_{i,1}\right] \leq \left\|\pi_i(.|X_0^{-i}) - \pi_i(.|Y_0^{-i})\right\|_{TV} \tag{1}$$

2.1 Modeling Asynchronicity

We use the asynchronicity model from [RRWN11] and [DSOR16]. Hogwild!-Gibbs is a multi-threaded algorithm where each thread performs a Gibbs update on the state of a graph which is stored in shared memory (typically in RAM). We view each processor's write as occuring at a distinct time instant. And each write starts the next time step for the process. Assuming that the writes are all serialized, one can now talk about the state of the system after t writes. This will be denoted as time t. HOGWILD! is modeled as a stochastic system adapted to a natural filtration $\mathcal{F}_t$. $\mathcal{F}_t$ contains all events thast have occured until time t. Some of these writes happen based on a read done a few steps ago and hence correspond to updates based on stale values in the local cache of the processor. The staleness is modeled in a stochastic manner using the random variable $\tau_{i,t}$ to denote the delay associated with the read performed on node i at time step t. The value of node i used in the update at time t is going to be $Y_{i,t} = X_{i,(t-\tau_{i,t})}$. Delays across different node reads can be correlated. However delay distribution is independent of the configuration of the model at time t. The model imposes two restrictions on the delay distributions. First, the expected value of each delay distribution is bounded by τ. We will think of τ as a constant compared to n in this paper. We call τ the average contention parameter associated with a HOGWILD!-Gibbs execution. [DSOR16] impose a second restriction which bounds the tails of the distribution of $\tau_{i,t}$. We do not need to make this assumption in this paper for our results. [DSOR16] need the assumption to show that the HOGWILD! chain mixes fast. However, by using coupling arguments we can avoid the need to have the HOGWILD! chain mix and will just use the mixing time bounds for the sequential Gibbs sampling chain instead. Let T denote the set of all delay distributions. We refer to the sequential Gibbs sampler associated with a distribution π as G_π and the HOGWILD! Gibbs sampler together with T associated with a distribution p by H_p^{T}. Note that H_π is a time-inhomogenuous Markov chain and might not converge to a stationary distribution.

2.2 Properties of Polynomials on Ising Models satisfying Dobrushin's condition

Here we state some known results about polynomial functions on Ising models satisfying Dobrushin's condition.

Theorem 2 (Concentration of Measure for Polynomial Functions of the Ising model, [DDK17, GLP17, GSS18])**.** *Consider an Ising model p without external field on a graph $G = (V, E)$ satisfying Dobrushin's condition with Dobrushin parameter $\alpha < 1$. Let f_a be a degree d-polynomial over the Ising model. Let $X \sim p$. Then, there is a constant $c(\alpha, \delta)$, such that,*

$$\Pr\left[|f_a(X) - \mathbf{E}\left[f_a(X)\right]| > t\right] \leq 2\exp\left(-\frac{(1-\alpha)t^{2/d}}{c(\alpha,d)\left\|a\right\|_\infty^{2/d} n}\right).$$

As a corollary this also implies,

$$\mathbf{Var}\left[f_a(X)\right] \leq C_3(d,\alpha)n^d.$$

Theorem 3 (Concentration of Measure for Polynomial Functions of the Ising model, [DDK17, GLP17, GSS18])**.** *Consider an Ising model p without external field on a graph $G = (V, E)$ satisfying Dobrushin's condition with Dobrushin parameter $\alpha < 1$. Let f_a be a degree d-polynomial over the Ising model. Let $X \sim p$. Then, there is a constant $c(\alpha, \delta)$, such that,*

$$\Pr\left[|f_a(X) - \mathbf{E}\left[f_a(X)\right]| > t\right] \leq 2\exp\left(-\frac{(1-\alpha)t^{2/d}}{c(\alpha,d)\left\|a\right\|_\infty^{2/d} n}\right).$$

As a corollary this also implies,

$$\mathbf{Var}\left[f_a(X)\right] \leq C_3(d,\alpha)n^d.$$

Theorem 4 (Marginals Bound under Dobrushin's condition, [DDK17])**.** *Consider an Ising model p satisfying Dobrushin's condition with Dobrushin parameter $\alpha < 1$. For some positive integer d, let $f_a(x)$ be a degree d polynomial function. Then we have that, if $X \sim p$,*

$$\left|\mathbf{E}\left[f_a(X)\right]\right| \leq 2\left(\frac{4nd\log n}{1-\alpha}\right)^{d/2}.$$

3 Bounding The Expected Hamming Distance Between Coupled Executions of HOGWILD! and Sequential Gibbs Samplers

In this Section, we show that under the greedy coupling of the sequential and asynchronous chains, the expected Hamming distance between the two chains at any time t is small. This will form the basis for our accurate estimation results of Section 4. We begin by showing that the expected Hamming distance between the states X_t and Y_t of a coupled run of the sequential and asynchronous executions respectively, is bounded by a $(\tau\alpha\log n)/(1-\alpha)$. At a high level, the proof of Lemma 2 proceeds by studying the expected change in the Hamming distance under one step of the coupled execution of the chains. We can bound the expected change using the Dobrushin parameter and the property of the greedy coupling (Lemma 1). We then show that the expected change is negative whenever the Hamming distance between the two chains was above $O(\log n)$ to begin with. This allows us to argue that when the two chains start at the same configuration, then the expected Hamming distance remains bounded by $O(\log n)$.

Lemma 2. *Let π denote a discrete probability distribution on n variables (nodes) with Dobrushin parameter $\alpha < 1$. Let $G_\pi = X_0, X_1, \ldots, X_t, \ldots$ denote the execution of the sequential Gibbs sampler on π and $H_\pi^{\mathrm{T}} = Y_0, Y_1, \ldots, Y_t, \ldots$ denote the HOGWILD! Gibbs sampler associated with π such that $X_0 = Y_0$. Suppose the two chains are running coupled in a greedy manner. Let $\mathcal{K}_t$ denote all events that have occured until time t in this coupled execution. Then we have, for all $t \geq 0$, under the greedy coupling of the two chains,*

$$\mathbf{E}\left[d_H(X_t, Y_t)|\mathcal{K}_0\right] \leq \frac{\tau\alpha\log n}{1-\alpha}$$

Proof. We will show the statement of the Lemma is true by induction over t. The statement is clearly true at time $t = 0$ since $X_0 = Y_0$. Suppose it holds for some time $t > 0$. The state of the system at time t, $\mathcal{K}_t$ includes the choice of nodes the two chains chose to update at each time step, the delays $\tau_{i,t'}$ for each node i and time step $t' \leq t$ and the states $\{X_{t'}\}_{t \geq t' \geq 0}$ and $\{Y_{t'}\}_{t \geq t' \geq 0}$ under the two chains. We let $I_{t'}$ denote the node that was chosen to be updated in the two chains at time step t'. As a shorthand, we use L_t to denote $d_H(X_t, Y_t)$. We will first compute a bound on

$$\mathbf{E}\left[L_{t+1}|\mathcal{K}_t\right] \tag{2}$$

The induction hypothesis implies that $\mathbf{E}[L_t|\mathcal{K}_0] \leq \frac{\tau\alpha}{1-\alpha}$. Given the delays for time $t+1$, denote by Y_t' the following state

$$Y_{i,t}^{\tau} = Y_{i,t-\tau_{i,t+1}} \; \forall \, i.$$

We partition the set of nodes into the set where X_t and Y_t have the same value and the set where they don't. Define $V_t^{=}$ and $V_t^{\neq}$ as follows.

$$V_t^{=} = \{i \in [n] \; s.t. \; X_{i,t} = Y_{i,t}\}$$
$$V_t^{\neq} = \{i \in [n] s.t. X_{i,t} \neq Y_{i,t}\}.$$

When the two samplers proceed to perform their update for time step $t+1$, in addition to the nodes in the set $V_t^{\neq}$ some additional nodes might appear to have different values under the asynchronous chain.

This is because of the delays $\tau_{i,t+1}$. We use D_{t+1} to denote the set of nodes in $V_t^=$ whose values read by the asynchronous chain are different from those under the sequential chain.

$$D_{t+1} = \left\{ i \in [n] \;\middle|\; X_{i,t} \neq Y_{i,t-\tau_{i,t+1}} \right\} \tag{3}$$

Next, we proceed to obtain a bound on the expected size of D_{t+1} which we will use later. Let $\tau_{1,t}, \tau_{2,t}, \ldots, \tau_{n,t}$ denote the delays at time t. Let $\delta_{i,t}$ denote the last index before t when the node i's value was updated. Observe that the $\delta_{i,t}$s have to all be distinct. Then

$$\mathbf{E}\left[|D_{t+1}||\mathcal{K}_t\right] \leq \sum_{i \in V_t^=} \Pr\left[\tau_{i,t} > \delta_{i,t}\right] \leq \sum_{i \in V_t^=} \frac{\tau}{\delta_{i,t}} \leq \tau \left(\sum_{j=1}^{|V_t^=|} \frac{1}{i} \right) \leq \tau \log n. \tag{4}$$

Suppose a node i from the set $V_t^=$ was chosen at step $t+1$ to be updated in both the chains. Now, L_{t+1} is either L_t or $L_t + 1$. The probability that it is $L_t + 1$ is bounded above by the total variation between the corresponding conditional distributions.

$$\Pr\left[X_{i,t+1} \neq Y_{i,t+1} \mid \mathcal{K}_t,\ i \in V_t^= \text{ chosen in step } t+1\right] \leq \left\| \pi_i(.|X_t^{-i}) - \pi_i(.|Y_t^{\tau^{-i}}) \right\|_{TV} \tag{5}$$

$$\leq \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i). \tag{6}$$

where (5) follows from the property of the greedy coupling (Lemma 1) and (6) follows from the triangle inequality because total variation distance is a metric.

Now suppose that i was chosen instead from the set $V_t^{\neq}$. Now L_{t+1} is either L_t or $L_t - 1$. We lower bound the probability that it is $L_t - 1$ using arguments similar to the above calculation.

$$\Pr\left[X_{i,t+1} = Y_{i,t+1} \middle| \mathcal{K}_t,\ i \in V_t^{\neq} \text{ was chosen in step } t+1\right] \geq 1 - \left\| \pi_i(.|X_t^{-i}) - \pi_i(.|Y_t^{\tau^{-i}}) \right\|_{TV} \tag{7}$$

$$\geq 1 - \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i). \tag{8}$$

where (7) follows from the property of the greedy coupling (Lemma 1) and (8) follows from the triangle inequality because total variation distance is a metric.

Now the expected change in the Hamming distance

$$\mathbf{E}\left[L_{t+1} - L_t | \mathcal{K}_t\right] = \frac{1}{n} \mathbf{E}\left[\sum_{i \in V_t^=} \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i) - \sum_{i \in V_t^{\neq}} \left(1 - \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i) \right) | \mathcal{K}_t \right]$$

$$\leq \frac{1}{n} \mathbf{E}\left[\sum_{j \in V_t^{\neq} \cup D_{t+1}} \sum_{i \in V} I(j,i) - L_t \middle| \mathcal{K}_t \right] \leq \frac{\left(L_t + \mathbf{E}\left[|D_{t+1}||\mathcal{K}_t\right]\right)\alpha}{n} \leq \frac{(L_t + \tau \log n)\alpha - L_t}{n}.$$

where we used the fact that $\sum_{i \in V} I(j,i) \leq \alpha$ for any j.

Now,

$$\mathbf{E}\left[L_{t+1} | \mathcal{K}_0\right] = \mathbf{E}\left[\mathbf{E}\left[L_{t+1} | \mathcal{K}_t\right] | \mathcal{K}_0\right] \tag{9}$$

$$\leq \mathbf{E}\left[L_t + \frac{(L_t + \tau \log n)\alpha - L_t}{n} \middle| \mathcal{K}_0 \right] \leq \frac{\tau\alpha \log n}{1-\alpha} \left(1 - \frac{1-\alpha}{n} \right) + \frac{\tau\alpha \log n}{n} \tag{10}$$

$$= \frac{\tau\alpha \log n}{1-\alpha}. \tag{11}$$

□

Next, we generalize the above Lemma to bound also the d^{th} moment of the Hamming distance between X_t and Y_t obtained from the coupled executions. The proof of Lemma 3 follows a similar flavor as that of Lemma 2. It is however more involved to bound the expected increase in the d^{th} power of the Hamming distance and it requires some careful analysis to see that the bound doesn't scale polynomially in n.

Lemma 3 (d^{th} moment bound on Hamming)**.** *Consider the same setting as that of Lemma 2. We have, for all $t \geq 0$, under the greedy coupling of the two chains,*

$$\mathbf{E}\left[d_H(X_t, Y_t)^d | \mathcal{K}_0\right] \leq C(\tau, \alpha, d) \log^d n,$$

where $C(.)$ is some function of the parameters τ, α and d.

Proof. Again we will employ the shorthand $L_t = d_H(X_t, Y_t)$. The proof is structured in the following way. We will show the statement of the Lemma is true by induction over t and d. To show the statement for a certain values of t, d we will use the inductive hypotheses of the statement for all t', d where $t' < t$ and for all t, d' where $d' < d$. Lemma 2 shows the statement for all values of t for $d = 1$. We will assume the statement holds for all t for $d' < d$.

Now, we proceed to show it is true for d. Clearly for $t = 0$ it holds because $d_H(X_0, Y_0)^d = 0$. Suppose it holds for some time $t > 0$. The state of the system at time t, $\mathcal{K}_t$ includes the choices of nodes the two chains chose to update at each time step, the delays $\tau_{i,t'}$ for each node i and time step $t' \leq t$ and the states $\{X_{t'}\}_{t \geq t' \geq 0}$ and $\{Y_{t'}\}_{t \geq t' \geq 0}$ under the two chains. We let $I_{t'}$ denote the node that was chosen to be updated in the two chains at time step t'. We will proceed in a similar way as we did in Lemma 2. We first compute a bound on

$$\mathbf{E}\left[L_{t+1}^d | \mathcal{K}_t\right] \tag{12}$$

as a function of L_t. The induction hypothesis implies that $\mathbf{E}[L_t^d | \mathcal{K}_0] \leq C(\tau, \alpha, d) \log^d n$. We partition the set of nodes into the set where X_t and Y_t have the same value and the set where they don't. Define $V_t^{=}$ and $V_t^{\neq}$ as follows.

$$V_t^{=} = \{i \in [n] \; s.t. \; X_{i,t} = Y_{i,t}\}$$
$$V_t^{\neq} = \{i \in [n] s.t. X_{i,t} \neq Y_{i,t}\}.$$

When the two samplers proceed to perform their update for time step $t+1$, in addition to the nodes in the set $V_t^{\neq}$ some additional nodes might appear to have different values under the asynchronous chain. This is because of the delays $\tau_{i,t+1}$. We use D_{t+1} to denote the set of nodes in $V_t^{=}$ whose values read by the asynchronous chain are different from those under the sequential chain.

$$D_{t+1} = \left\{i \in [n] \;\middle|\; X_{i,t} \neq Y_{i,t-\tau_{i,t+1}}\right\} \tag{13}$$

Next, we use the bound on the expected size of D_{t+1} which was derived in Lemma 2.

$$\mathbf{E}\left[|D_{t+1}| | \mathcal{K}_t\right] \leq \tau \log n. \tag{14}$$

Suppose a node i from the set $V_t^{=}$ was chosen at step $t+1$ to be updated in both the chains. Now, L_{t+1} is either L_t or $L_t + 1$. The probability that it is $L_t + 1$ is bounded above by the total variation between the corresponding conditional distributions.

$$\Pr\left[X_{i,t+1} \neq Y_{i,t+1} | \mathcal{K}_t, \; i \in V_t^{=} \text{ chosen in step } t+1\right] \leq \left\|\pi_i(.|X_t^{-i}) - \pi_i(.|Y_t^{\tau^{-i}})\right\|_{TV} \tag{15}$$
$$\leq \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j, i). \tag{16}$$

(16) follows again due to the property of greedy coupling (Lemma 1) and the metric property of the total variation distance.

Now suppose that i was chosen instead from the set $V_t^{\neq}$. Now L_{t+1} is either L_t or $L_t - 1$. The probability that it is $L_t - 1$ is bounded below by the following.

$$\Pr\left[X_{i,t+1} = Y_{i,t+1} \middle| \mathcal{K}_t,\ i \in V_t^{\neq} \text{ chosen in step } t+1\right] \geq 1 - \left\|\pi_i(.|X_t^{-i}) - \pi_i(.|Y_t^{\tau^{-i}})\right\|_{TV} \tag{17}$$

$$\geq 1 - \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i). \tag{18}$$

(16) follows again due to the property of greedy coupling (Lemma 1) and the metric property of the total variation distance.

Now the expected change in the value of the d^{th} power of the Hamming distance is

$$\mathbf{E}\left[L_{t+1}^d - L_t^d | \mathcal{K}_t\right] \tag{19}$$

$$= \frac{1}{n}\mathbf{E}\left[\sum_{i \in V_t^{=}} \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i)\left((L_t+1)^d - L_t^d\right) - \sum_{i \in V_t^{\neq}}\left(1 - \sum_{j \in V_t^{\neq} \cup D_{t+1}} I(j,i)\right)\left(L_t^d - (L_t-1)^d\right)|\mathcal{K}_t\right] \tag{20}$$

$$\leq \frac{1}{n}\left((L_t+1)^d - L_t^d\right)\mathbf{E}\left[\sum_{j \in V_t^{\neq} \cup D_{t+1}} \sum_{i \in V} I(j,i)\right] - \frac{L_t}{n}\left(L_t^d - (L_t-1)^d\right) \tag{21}$$

$$\leq \frac{1}{n}\left((L_t+1)^d - L_t^d\right)\left(L_t + \mathbf{E}\left[D_{t+1}|\mathcal{K}_t\right]\right)\alpha - \frac{L_t}{n}\left(L_t^d - (L_t-1)^d\right) \tag{22}$$

$$\leq \frac{1}{n}\left((L_t+1)^d - L_t^d\right)\left(L_t + \tau \log n\right)\alpha - \frac{L_t}{n}\left(L_t^d - (L_t-1)^d\right) \tag{23}$$

Now, suppose $\mathbf{E}\left[L_t^d | \mathcal{K}_0\right] \leq C(\tau,\alpha,d)\log^d n - c_3(d)C(\tau,\alpha,d-1)\log^{d-1} n$ for an appropriate constant $c_3(d)$.Then,

$$\mathbf{E}\left[L_{t+1}^d - L_t^d | \mathcal{K}_0\right] = \mathbf{E}\left[(L_{t+1} - L_t)\left(\sum_{i=0}^{d-1} L_{t+1}^d - i - 1 L_t^i\right)\right] \tag{24}$$

$$\leq \mathbf{E}\left[\sum_{i=0}^{d-1} L_{t+1}^{d-i-1} L_t^i\right] \leq c_3(d)C(\tau,\alpha,d-1)\log^{d-1} n \tag{25}$$

$$\implies \mathbf{E}\left[L_{t+1}^d\right] \leq C(\tau,\alpha,d)\log^d n, \tag{26}$$

where (25) follows because

$$\mathbf{E}\left[L_{t+1}^{d-i-1} L_t^i\right] \leq \mathbf{E}\left[L_t^i (L_t+1)^{d-i-1}\right] \leq \mathbf{E}\left[L_t^{d-1}\right] + c(d)o\left(\mathbf{E}\left[L_t^{d-1}\right]\right). \tag{27}$$

Suppose instead that $\mathbf{E}\left[L_t^d | \mathcal{K}_0\right]$ was larger than $C(\tau,\alpha,d)\log^d n - c_3(d)C(\tau,\alpha,d-1)\log^{d-1} n$. We want to show that for $C(\tau,\alpha,d)$ appropriately large in d, $\mathbf{E}\left[L_{t+1}^d | \mathcal{K}_0\right]$ doesn't exceed $C(\tau,\alpha,d)$.

$$\mathbf{E}\left[L_{t+1}^d | \mathcal{K}_0\right] = \mathbf{E}\left[\mathbf{E}\left[L_{t+1}^d | \mathcal{K}_t\right] | \mathcal{K}_0\right] \tag{28}$$

$$\leq \mathbf{E}\left[L_t^d + \frac{(L_t + \tau\log n)\alpha}{n}\left((L_t+1)^d - L_t^d\right) - \frac{L_t}{n}\left(L_t^d - (L_t-1)^d\right)\middle|\mathcal{K}_0\right] \tag{29}$$

$$= \mathbf{E}\left[L_t^d + \frac{(L_t + \tau\log n)\alpha}{n}\left(\sum_{i=1}^{d}\binom{d}{i} L_t^{d-i}\right) - \frac{L_t}{n}\left(\sum_{i=1}^{d}\binom{d}{i} L_t^{d-i}(-1)^i\right)\middle|\mathcal{K}_0\right] \tag{30}$$

$$\leq C(\tau,\alpha,d)\log^d n - \frac{1}{n}\left(\mathbf{E}[L_t^d](1-\alpha) - \sum_{i=2}^{d}\binom{d}{i} L_t^{d-i+1}(1+\alpha) - \sum_{i=1}^{d}\binom{d}{i} L_t^{d-i}\tau\alpha\log n\right) \tag{31}$$

$$\leq C(\tau,\alpha,d)\log^d n, \tag{32}$$

where (32) holds because for $C(\tau,\alpha,d)$ sufficiently large compared to the values of $C(\tau,\alpha,d')$, where $d' < d$, we can have $C(\tau,\alpha,d)\log^d n - c_3(d)C(\tau,\alpha,d-1)\log^{d-1} n$ dominating the all the remaining terms in the two summations. Hence, by induction we have the desired Lemma statement. □

4 Estimating Global Functions Using HOGWILD! Gibbs Sampling

To begin with, we observe that our Hamming moment bounds from Section 3 imply that we can accurately estimate functions or events of the graphical model if they are Lipschitz. We show this below as a Corollary of Lemma 3. Before we state the Corollary, we will first state the following simple Lemma which quantifies how large a t is required to have an accurate estimate from the Gibbs sampler.

Lemma 4. *Let π be an graphical model on n nodes satisfying Dorbushin's condition with Dobrushin parameter $\alpha < 1$. Let $X_0, X_1, \ldots, X_t$ denote the steps of a Gibbs sampler running on π. Let $X \sim \pi$. Also let $f(x)$ be a bounded function on the graphical model. Then for $t > 0$,*

$$|\mathbf{E}[f_a(X_t)] - \mathbf{E}[f_a(X)]| \leq \|f\|_\infty n \exp\left(-\frac{(1-\alpha)t}{n}\right).$$

Proof. We have from Theorem 1 that $d_{\mathrm{TV}}(X_t, X) \leq n \exp\left(\frac{(1-\alpha)t}{n}\right)$. This implies

$$|\mathbf{E}[f_a(X_t)] - \mathbf{E}[f_a(X)]| \leq \left|\max_x f_a(x)\right| n \exp\left(\frac{(1-\alpha)t}{n}\right) \leq \|a\|_\infty n^{d+1} \exp\left(-\frac{(1-\alpha)t}{n}\right). \tag{33}$$

□

Now, we state Corollary 1 which quantifies the error we can attain when trying to estimate expectations of Lipschitz functions using HOGWILD!-Gibbs.

Corollary 1. *Let π denote the distribution associated with a graphical model over the set of variables V ($|V| = n$) taking values in a discrete space S^n. Assume that the model satisfies Dobrushin's condition with Dobrushin parameter $\alpha < 1$. Let $f : S^{|V|} \to \mathbb{R}$ be a function such that, for all $x, y \in S^{|V|}$,*

$$|f(x) - f(y)| \leq K d_H(x, y)^d.$$

Let $X \sim \pi$ and let $Y_0, Y_1, \ldots, Y_t$ denote an execution of HOGWILD!-Gibbs sampling on π with average contention parameter τ. For $t > \frac{n}{1-\alpha} \log(2\|f\|_\infty n/K)$,

$$|\mathbf{E}[f(Y_t)] - \mathbf{E}[f(X)]| \leq K.(C(\tau, \alpha, d) \log^d n + 1).$$

where $C(.)$ is the function from Lemma 3.

Proof. Consider an execution $X_0, X_1, \ldots, X_t$, where $X_0 = Y_0$, of the synchronous Gibbs sampling associated with π coupled greedily with the HOGWILD! chain. We have from 1 and 4 that, for $t > \frac{n}{1-\alpha} \log(2\|f\|_\infty n/K)$,

$$|\mathbf{E}[X_t] - \mathbf{E}[X]| \leq K. \tag{34}$$

Next, we have,

$$\mathbf{E}[f(X_t) - f(Y_t)] \leq \mathbf{E}\left[K d_H(X_t, Y_t)^d\right] \leq K.C(\tau, \alpha, d) \log^d n. \tag{35}$$

Putting (34) and (35) together we get the statement of the Corollary. □

We note that the results of [DSOR16] can be used to obtain Corollary 1 when the function is Lipschitz with respect to the Hamming distance. The above corollary provides a simple way to bound the bias introduced by HOGWILD! in estimation of Lipschitz functions. However, many functions of interest over graphical models are not Lipschitz with good Lipschitz constants. In many cases, even when the Lipschitz constants are bad, there is still hope for more accurate estimation. As it turns out Dobrushin's condition provides such cases. We will focus on one such case which is polynomial functions of the Ising model. Our goal will be to accurately estimate the expected values of constant degree polynomials over the Ising model. Using the bounds from Lemmas 2 and 3, we now proceed to bound the bias in computing polynomial functions of the Ising model using HOGWILD! Gibbs sampling.

We first remark that linear functions (degree 1 polynomials) suffer 0 bias in their expected values due to HOGWILD!-Gibbs. This is because under zero external field Ising models $\mathbf{E}[\sum_i a_i X_i] = 0$ since each node individually has equal probability of being ± 1. This symmetry is maintained by HOGWILD!-Gibbs since the delays are configuration-agnostic. Hence the delays when a node is $+1$ and when it is -1 can be coupled perfectly leaving the symmetry intact. More interesting cases start happening when we go to degree 2 polynomials. Therefore, we start our investigation at quadratic polynomials. Theorem 5 states the bound we show for the bias in computation of degree 2 polynomials of the Ising model.

Theorem 5 (Bias in Quadratic functions of Ising Model computed using HOGWILD!-Gibbs)**.** *Consider the quadratic function $f_a(x) = \sum_{i,j:i<j} a_{ij} x_i x_j$. Let p denote an Ising model on n nodes with Dobrushin parameter $\alpha < 1$. Let $\{X_t\}_{t\geq 0}$ denote a run of sequential Gibbs sampler and $H_p^{\mathrm{T}} = \{Y_t\}_{t\geq 0}$ denote a run of HOGWILD!- Gibbs on p, such that $X_0 = Y_0$. Then we have, for $t > \frac{6n}{1-\alpha}\log(2\|a\|_\infty n)$, under the greedy coupling of the two chains,*

$$|\mathbf{E}[f_a(X_t) - f_a(Y_t)]| \leq c_2 \|a\|_\infty \frac{\tau\alpha \log n}{(1-\alpha)^{3/2}} (n \log n)^{1/2}.$$

Proof. Under the greedy coupling, we have that,

$$|\mathbf{E}[f_a(X_t) - f_a(Y_t)]| = \tag{36}$$

$$\left| \mathbf{E}\left[\sum_{i, X_{i,t}\neq Y_{i,t}} \sum_{j>i, X_{j,t}=Y_{j,t}} a_{ij}\left(X_{i,t}X_{j,t} - Y_{i,t}Y_{j,t}\right)\right] + \mathbf{E}\left[\sum_{i, X_{i,t}=Y_{i,t}} \sum_{j>i, X_{j,t}\neq Y_{j,t}} a_{ij}\left(X_{i,t}X_{j,t} - Y_{i,t}Y_{j,t}\right)\right]\right| \tag{37}$$

$$= \left| \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_{j>i, X_{j,t}=Y_{j,t}} a_{ij} X_{j,t}\right] + \mathbf{E}\left[\sum_{i, X_{i,t}=Y_{i,t}} X_{i,t} \sum_{j>i, X_{j,t}\neq Y_{j,t}} a_{ij}\left(X_{j,t} - Y_{j,t}\right)\right]\right| \tag{38}$$

$$= \left| \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_{j>i} a_{ij} X_{j,t}\right] - \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_{j>i, X_{j,t}\neq Y_{j,t}} a_{ij} X_{j,t}\right]\right.$$
$$\left. + \mathbf{E}\left[\sum_i X_{i,t} \sum_{j>i} a_{ij}\left(X_{j,t} - Y_{j,t}\right)\right] - \mathbf{E}\left[\sum_{i, X_{i,t}\neq Y_{i,t}} X_{i,t} \sum_{j>i} a_{ij}\left(X_{j,t} - Y_{j,t}\right)\right]\right| \tag{39}$$

$$= \left| \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_j a_{ij} X_{j,t}\right] + \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_{j, X_{j,t}\neq Y_{j,t}} X_{j,t}\right]\right| \tag{40}$$

$$\leq \left| \mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_j a_{ij} X_{j,t}\right]\right| + \mathbf{E}\left[d_H(X_t, Y_t)^2\right]. \tag{41}$$

where (37) is based on the observation that if $X_{i,t}X_{j,t} = Y_{i,t}Y_{j,t}$ then the difference associated with this monomial vanishes, (39) and (40) follow via rearrangement of terms, and (41) follows because $\left|\sum_{j, X_{j,t}\neq Y_{j,t}} X_{j,t}\right| \leq d_H(X_t, Y_t)$.

We bound each term in (41) seperately. First let us consider the term $\left|\mathbf{E}\left[\sum_i (X_{i,t} - Y_{i,t}) \sum_j a_{ij} X_{j,t}\right]\right|$. We will employ concentration of measure of linear functions of the Ising model to bound this term. Intuitively when t is large enough, X_t is very close to a sample from the true Ising model and hence X_t will have the properties of a true sample from the Ising model. In particular, we can employ Theorem 3 to argue that if

$X \sim p$, then

$$\Pr\left[\left|\sum_j a_{ij}X_j\right| > r\right] \leq 2\exp\left(-\frac{r^2(1-\alpha)}{8\left\|a\right\|_\infty^2 n}\right) \ \forall\, i$$

$$\implies \Pr\left[\left|\sum_j a_{ij}X_j\right| \leq c\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}}\,\forall\, i\right] \geq 1-\frac{1}{n^3} \tag{42}$$

$$\implies \Pr\left[\left|\sum_j a_{ij}X_{j,t}\right| \leq c\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}}\,\forall\, i\right] \geq 1-\frac{2}{n^3} \tag{43}$$

where (42) holds for a large enough constant c and (43) holds via an application of Lemma 4 for bilinear functions of the Ising model on the fact that $t > \frac{6n}{1-\alpha}\log(2\left\|a\right\|_\infty n)$. Denote by the set G_i, the following set of states

$$G_i = \left\{x \in \Omega \middle| \left|\sum_j a_{ij}x_j\right| \leq c\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}}\right\} \tag{44}$$

Now,

$$\mathbf{E}\left[\sum_i (X_{i,t}-Y_{i,t})\sum_j a_{ij}X_{j,t}\right] \tag{45}$$

$$\leq \mathbf{E}\left[\sum_i (X_{i,t}-Y_{i,t})(c\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}})\middle|\, \forall\, i, X_t \in G_i\right] + \left\|a\right\|_\infty n^2 \Pr\left[\exists\, i:\ X_t \notin G_i\right] \tag{46}$$

$$\leq c\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}}\mathbf{E}\left[d_H(X_t,Y_t)|\forall\, i\ X_t \in G_i\right] + \frac{2}{n^3}. \tag{47}$$

where in (46) we used the fact that $\sum_i (X_i - Y_i)\sum_j a_{ij}X_j \leq \left\|a\right\|_\infty n^2$ and (47) follows because . Now,

$$\mathbf{E}\left[d_H(X_t,Y_t)|\forall\, i\ X_t \in G_i\right] \leq \frac{\mathbf{E}\left[d_H(X_t,Y_t)\right]}{\Pr\left[\forall\, i\ X_t \in G_i\right]} \leq 2\mathbf{E}\left[d_H(X_t,Y_t)\right] \leq \frac{2\tau\alpha\log n}{1-\alpha}, \tag{48}$$

where we have used that $\Pr[\forall i X_t \in G_i] \geq 1-2/n^3 \geq 1/2$ and employed Lemma 2. Hence, $\left|\mathbf{E}\left[\sum_i (X_{i,t}-Y_{i,t})\sum_j a_{ij}X_{j,t}\right]\right| \leq (c+1)\left\|a\right\|_\infty \sqrt{\frac{n\log n}{1-\alpha}}\frac{2\tau\alpha\log n}{1-\alpha}$. The second term we need to bound is

$$\mathbf{E}\left[d_H(X_t,Y_t)^2\right] \leq C(\tau,\alpha,2)\log^2 n \tag{49}$$

which follows from Lemma 3. Putting the two bounds together we get the statement of the Theorem. □

The main intuition behind the proof is that we can improve upon the bound implied by the Lipschitz constant by appealing to strong concentration of measure results about functions of graphical models under Dobrushin's condition [DDK17, GLP17, GSS18].

We extend the ideas in the above proof to bound the bias introduced by the HOGWILD!-Gibbs algorithm when computing the expected values of a degree d polynomial of the Ising model in high temperature. Our main result concerning d-linear functions is Theorem 6.

Theorem 6 (Bias in degree d polynomials computed using HOGWILD!-Gibbs)**.** *Consider a degree d polynomial of the form $f_a(x) = \sum_{i_1 i_2,\ldots,i_d} a_{i_1 i_2 \ldots i_d} x_{i_1} x_{i_2} \ldots x_{i_d}$. Consider the same setting as that of Theorem 5. Then we have, for $t > \frac{n(d+1)}{1-\alpha}\log n$, under the greedy coupling of the two chains,*

$$|\mathbf{E}[f_a(X_t) - f_a(Y_t)]| \leq c'\left\|a\right\|_\infty (n\log n)^{(d-1)/2}.$$

To show Theorem 6, we will use the following helper Lemmas: 5 and 6. We will state them first.
For simplicity, here we consider degree d polynomials of the form $f_a(x) = \sum_{i_1 i_2,\ldots,i_d} a_{i_1 i_2 \ldots i_d} x_{i_1} x_{i_2} \ldots x_{i_d}$.

Lemma 5. *Consider a degree d polynomial $f_a(x) = \sum_{i_1 i_2,\ldots,i_d} a_{i_1 i_2 \ldots i_d} x_{i_1} x_{i_2} \ldots x_{i_d}$. Let p denote a high temperature Ising model on a graph $G = (V, E)$ with $|V| = n$ nodes with Dobrushin parameter $\alpha < 1$. Let $G_p = X_0, X_1, \ldots, X_t, \ldots$ denote the synchronous Gibbs sampler on p and $H_p^{\mathrm{T}} = Y_0, Y_1, \ldots, Y_t, \ldots$ denote the HOGWILD! Gibbs sampler associated with p such that $X_0 = Y_0$. Then we have, for $t > \frac{n(d+1)}{1-\alpha} \log n$, under the greedy coupling of the two chains, for all $0 \leq k \leq d$,*

$$\left| \mathbf{E} \left[\sum_{i_1} (X_{i_1,t} - Y_{i_1,t}) \left(\ldots \left(\sum_{i_k} (X_{i_k,t} - Y_{i_k,t}) \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} X_{i_l} \right) \right) \right] \right|$$
$$\leq C(\tau, \alpha, k) \log^k n C_2(\alpha, d-k) (n \log n)^{(d-k)/2}.$$

Proof. We will employ the bound we have on the moments on the Hamming distance 3 together with concentration of measure for polynomial functions of the Ising model 3 to show the statement. We have from Theorems 3 and 4 and Lemma 4, that for every choice of $i_1, i_2, \ldots i_k \in V$, and for $t > \frac{n(d+1)}{1-\alpha} \log n$,

$$\Pr \left[\left| \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} X_{i_l,t} \right| > c_4(\alpha, d-k) \left(\frac{\|a\|_\infty n \log n}{1-\alpha} \right)^{(d-k)/2} \right] \leq \frac{1}{n^{d+2} \|a\|_\infty} \tag{50}$$

$$\implies \Pr \left[\exists \{i_1, i_2, \ldots i_k \in V\} \left| \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} X_{i_l,t} \right| > c_4(\alpha, d-k) \left(\frac{n \log n}{1-\alpha} \right)^{(d-k)/2} \right] \tag{51}$$

$$\leq \frac{1}{n^{d-k+2} \|a\|_\infty}, \tag{52}$$

where we used the fact that when t is large, the distribution of X_t is close to p (Lemma 4) and (51) follows from a union bound. Define the set of states $G_{i_1 i_2 \ldots i_k}$ as follows.

$$G_{i_1 i_2 \ldots i_k} = \left\{ x \in \Omega \;\middle|\; \left| \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} x_{i_l} \right| \leq c_4(\alpha, d-k) \left(\frac{n \log n}{1-\alpha} \right)^{(d-k)/2} \forall \{i_1, i_2, \ldots i_k \in V\} \right\} \tag{53}$$

Then we have,

$$\left| \mathbf{E} \left[\sum_{i_1} (X_{i_1,t} - Y_{i_1,t}) \left(\ldots \left(\sum_{i_k} (X_{i_k,t} - Y_{i_k,t}) \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} X_{i_l,t} \right) \right) \right] \right| \tag{54}$$

$$\leq \mathbf{E} \left[\sum_{i_1} |X_{i_1,t} - Y_{i_1,t}| \left(\ldots \left(\sum_{i_k} |X_{i_k,t} - Y_{i_k,t}| \left| \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \prod_{l=k+1}^{d} X_{i_l,t} \right| \right) \right) \right] \tag{55}$$

$$\leq c_4(\alpha, d-k) \left(\frac{n \log n}{1-\alpha} \right)^{(d-k)/2} \mathbf{E} \left[\left(\sum_i |X_{i,t} - Y_{i,t}| \right)^k \;\middle|\; X_t \in G_{i_1 \ldots i_k} \forall \{i_1, \ldots, i_k \in V\} \right] + \frac{1}{n^2} \tag{56}$$

$$\leq c_4(\alpha, d-k) \left(\frac{n \log n}{1-\alpha} \right)^{(d-k)/2} 2^k \mathbf{E} \left[d_H(X_t, Y_t)^k \;\middle|\; X_t \in G_{i_1 \ldots i_k} \forall \{i_1, \ldots, i_k \in V\} \right] + \frac{1}{n^2} \tag{57}$$

$$\leq c_4(\alpha, d-k) \left(\frac{n \log n}{1-\alpha} \right)^{(d-k)/2} 2^k . 2C(\tau, \alpha, k) \log^k n \leq C(\tau, \alpha, k) \log^k n C_2(\alpha, d-k) (n \log n)^{(d-k)/2}. \tag{58}$$

where we have used the fact that $\sum_i |X_i - Y_i| = 2 d_H(X, Y)$ for $X, Y \in \Omega$, and (58) follows from Lemma 3 and the observation that $\Pr [X_t \in G_{i_1 \ldots i_k} \forall \{i_1, \ldots, i_k \in V\}] \geq 1 - \frac{1}{n^{d-k+2} \|a\|_\infty} \geq 1/2$. □

Lemma 6. *Consider a degree d polynomial $f_a(x) = \sum_{i_1 i_2,\ldots,i_d} a_{i_1 i_2 \ldots i_d} x_{i_1} x_{i_2} \ldots x_{i_d}$. Let p denote a high temperature Ising model on a graph $G = (V, E)$ with $|V| = n$ nodes with Dobrushin parameter $\alpha < 1$. Let $G_p = X_0, X_1, \ldots, X_t, \ldots$ denote the synchronous Gibbs sampler on p and $H_p^{\mathrm{T}} = Y_0, Y_1, \ldots, Y_t, \ldots$ denote the HOGWILD! Gibbs sampler associated with p such that $X_0 = Y_0$. Then we have, for $t > \frac{n(d+1)}{1-\alpha} \log n$, under the greedy coupling of the two chains, for all $0 \leq k \leq d$,*

$$\left| \mathbf{E}\left[\sum_{i_1} (X_{i_1,t} - Y_{i_1,t}) \left(\ldots \left(\sum_{i_k} (X_{i_k,t} - Y_{i_k,t}) \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \left(\prod_{l=k+1}^{d} X_{i_l,t} - \prod_{l=k+1}^{d} Y_{i_l,t} \right) \right) \right) \right] \right|$$
$$\leq C(\tau, \alpha, k) \log^k n C_2(\alpha, d-k) (n \log n)^{(d-k)/2}.$$

Proof. We prove this by inducting backwards on k. When $k = d$, we get the desired statement from Lemma 3. Assume the statement holds for some $k + 1 < d$. We will show it holds for k as well. In the following calculations, we will, for a while, drop the subscript t from X_t and Y_t for brevity. We have,

$$\left| \mathbf{E}\left[\sum_{i_1} (X_{i_1} - Y_{i_1}) \left(\ldots \left(\sum_{i_k} (X_{i_k} - Y_{i_k}) \sum_{i_{k+1},\ldots,i_d} a_{i_1 \ldots i_d} \left(\prod_{l=k+1}^{d} X_{i_l} - \prod_{l=k+1}^{d} Y_{i_l} \right) \right) \right) \right] \right| \tag{59}$$
$$= \left| \mathbf{E}\left[\sum_{i_1} (X_{i_1} - Y_{i_1}) \left(\ldots \left(\sum_{i_k} (X_{i_k} - Y_{i_k}) \sum_{\substack{i_{k+1} \ldots i_d \\ X_{i_{k+1}} \ldots X_{i_d} \neq Y_{i_{k+1}} \ldots Y_{i_d}}} a_{i_1 \ldots i_d} \left(2 \prod_{l=k+1}^{d} X_{i_l} \right) \right) \right) \right] \right|. \tag{60}$$

The last summation of (60) is over indices $i_{k+1}, \ldots, i_d$ such that the product of values in X at these indices disagrees with the corresponding product of values in Y. This can happen due to a disagreement between X and Y at one of the indices $i_{k+1}, \ldots, i_d$, say j, and an agreement of the product over the rest of the indices. That is, $X_j \neq Y_j$ and $\prod_{l=k+1}^{j-1} X_{i_j} \prod_{l=j+1}^{d} X_{i_j} = \prod_{l=k+1}^{j-1} Y_{i_j} \prod_{l=j+1}^{d} Y_{i_j}$. This leads to the last summation in (60) being bounded above by the sum of $d - k$ terms of the form

$$\sum_{i_j} (X_{i_j} - Y_{i_j}) \sum_{i_{k+1},\ldots,i_{j-1},i_{j+1},\ldots,i_d} a_{i_1 \ldots i_d} \left(\prod_{l=k+1}^{j-1} X_{i_l} \prod_{l=j+1}^{d} X_{i_l} - \prod_{l=k+1}^{j-1} Y_{i_l} \prod_{l=j+1}^{d} Y_{i_l} \right). \tag{61}$$

Replacing the last summation of (60) with $d - k$ terms of the above form we see that the whole quantity decomposes into $d-k$ terms such that the inductive hypothesis can be applied over each of the terms. It is a fairly simple calculation to see that then we get the desired bound of the Theorem by induction (by keeping in mind that the quantity of focus here is the dependence on n and we treat d as a constant). □

Given Lemma 6, Theorem 6 follows. *Proof of Theorem 6:* The Theorem follows directly from Lemma 6 applied to the case $k = 0$. □

Next, we show that we can accurately estimate the expectations above by showing that the variance of the functions under the asynchronous model is comparable to that of the functions under the sequential model.

Theorem 7 (Variance of degree d polynomials computed using HOGWILD!-Gibbs)**.** *Consider a high temperature Ising model p on n nodes with Dobrushin parameter $\alpha < 1$. Let $f_a(x)$ be a degree d polynomial function Let $Y_0, Y_1, \ldots, Y_t$ denote a run of HOGWILD! Gibbs sampling associated with p. We have, for $t > \frac{(d+1)n}{1-\alpha} \log\left(n^2\right)$,*

$$\mathbf{Var}\left[f(Y_t)\right] \leq \|a\|_\infty^2 C(d, \alpha, \tau) n^d.$$

Proof.

$$\mathbf{Var}\left[f_a(Y_t)\right] = \mathbf{E}\left[f_a(Y_t)^2\right] - \mathbf{E}\left[f_a(Y_t)\right]^2 \tag{62}$$
$$\leq \mathbf{E}\left[f_a(Y_t)^2\right] + \mathbf{E}\left[f_a(Y_t)\right]^2 \tag{63}$$

First, we proceed to bound $\mathbf{E}\left[f_a(Y_t)^2\right]$. Consider a coupled execution of synchronous Gibbs sampling $X_0, X_1, \ldots, X_t$ where $X_0 = Y_0$ coupled using the greedy coupling. We have,

$$\mathbf{E}\left[f_a(Y_t)^2\right] \leq \mathbf{E}\left[f_a(X_t)^2\right] + \left|\mathbf{E}\left[f_a(X_t)^2 - f_a(Y_t)^2\right]\right| \tag{64}$$

$$\leq C_1(2d,\alpha)\left(n\log n\right)^d + C_2(2d,\alpha,\tau)\left(n\log n\right)^{(2d-1)/2} \leq C_3(d,\alpha,\tau)\left(n\log n\right)^d \tag{65}$$

where we used the fact that $f(x)^2$ is a degree $2d$ polynomial and then applied Theorem 6 for degree $2d$ polynomials, in addition to Theorem 4. Now we look at the second term of (63).

$$\mathbf{E}\left[f_a(Y_t)\right]^2 \leq \mathbf{E}\left[f_a(X_t)\right]^2 + \left|\mathbf{E}\left[f_a(Y_t)\right]^2 - \mathbf{E}\left[f_a(X_t)\right]^2\right| \tag{66}$$

$$\leq \mathbf{E}\left[f_a(X_t)\right]^2 + \left|\left(\mathbf{E}\left[f_a(Y_t)\right] - \mathbf{E}\left[f_a(X_t)\right]\right)\left(\mathbf{E}\left[f_a(Y_t)\right] + \mathbf{E}\left[f_a(X_t)\right]\right)\right| \tag{67}$$

$$\leq C_1^2(d,\alpha)\left(n\log n\right)^d + C_2(d,\alpha,\tau)\left(n\log n\right)^{(d-1)/2}\left(\left|2\mathbf{E}\left[f_a(X_t)\right]\right| + \left|\mathbf{E}\left[f_a(Y_t)\right] - \mathbf{E}\left[f_a(X_t)\right]\right|\right) \tag{68}$$

$$\leq C_4(d,\alpha,\tau)\left(n\log n\right)^{(2d-1)/2}, \tag{69}$$

where again we employ Theorem 4 and Theorem 6 for degree d polynomials. Combining (65) and (69), we get the desired bound. □

4.1 Going Beyond Ising Models

We presented results for accurate estimation of polynomial functions over the Ising model. However, the results can be extended to hold for more general graphical models satisfying Dobrushin's condition. A main ingredient here was concentration of measure. If the class of functions we look at has d^{th}-order bounded differences in expectation, then we indeed get concentration of measure for these functions (Theorem 1.2 of [GSS18]). This combined with the techniques in our paper would allow similar gains in accurate estimation of such functions on general graphical models.

5 Experiments

We show the results of experiments run on a machine with four 10-core Intel Xeon E7-4850 CPUs to demonstrate the practical validity of our theory. In our experiments, we focused on two Ising models—Curie-Weiss and the Grid. The Curie-Weiss $CW(n,\alpha)$ is the Ising model corresponding to the complete graph on n vertices with edges of weight $\beta = \frac{\alpha}{n-1}$. The $Grid(k^2,\alpha)$ model is the Ising model corresponding to the k-by-k grid with the left connected to the right and top connected to the bottom to form a torus—a four-regular graph; the edge weights are $\frac{\alpha}{4}$. The total influence of each of these models is at most α, so we chose $\alpha = 0.5$ to ensure Dobrushin's condition. To generate samples, we start at a uniformly random configuration and run Markov chains for $T = 10n\log_2(n)$ steps to ensure mixing.

In our first experiment (Figure 1) we validate the modeling assumption that the average delay of a read τ is a constant. Computing the exact delays in a real run of the HOGWILD! is not possible, but we approximate the delays by making processes log read and write operations to a lock-free queue as they execute the HOGWILD!-updates. We present two plots of the average delay of a read in a HOGWILD! run of the $CW(n,0.5)$ Markov chain with respect to n. Four asynchronous processors were used to generate the first plot, while twenty were used for the second. We notice that the average delay depends on the number of asynchronous processes, but is constant with respect to n as assumed in our model.

Next, we plot (in Figure 2) the relationship between the number of asynchronous processors used in a HOGWILD! execution and the delay parameter τ. For this plot, we estimated τ by the average empirical delay over HOGWILD! runs of $CW(n,0.5)$ models, with n ranging from 100 to 1000 in increments of one hundred. The plot shows a linear relationship, and suggests that the delay per additional processor is approximately 0.4 steps.

The primary purpose of our work is to demonstrate that polynomial statistics computed from samples of a HOGWILD! run of Gibbs Sampling will approximate those computed from a sequential run. Our third experiment demonstrates exactly this fact. We plot (in Figure 3 on the left) the empirical expectations of the

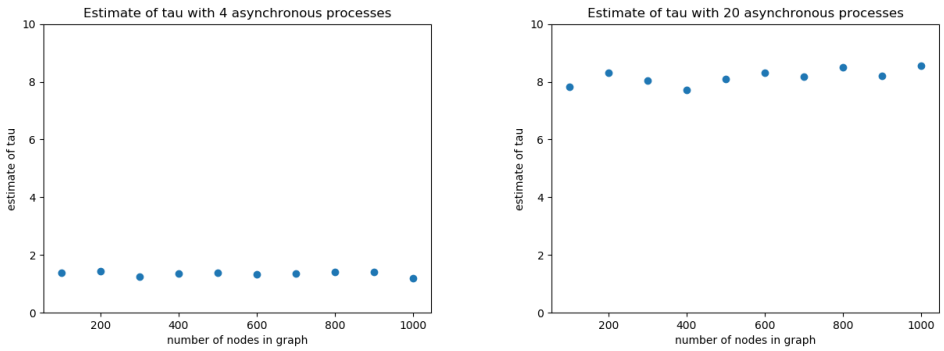

Figure 1: Average delay of reads for $CW(n, 0.5)$ model. Four asynchronous processors were used on the left, while twenty were used on the right.

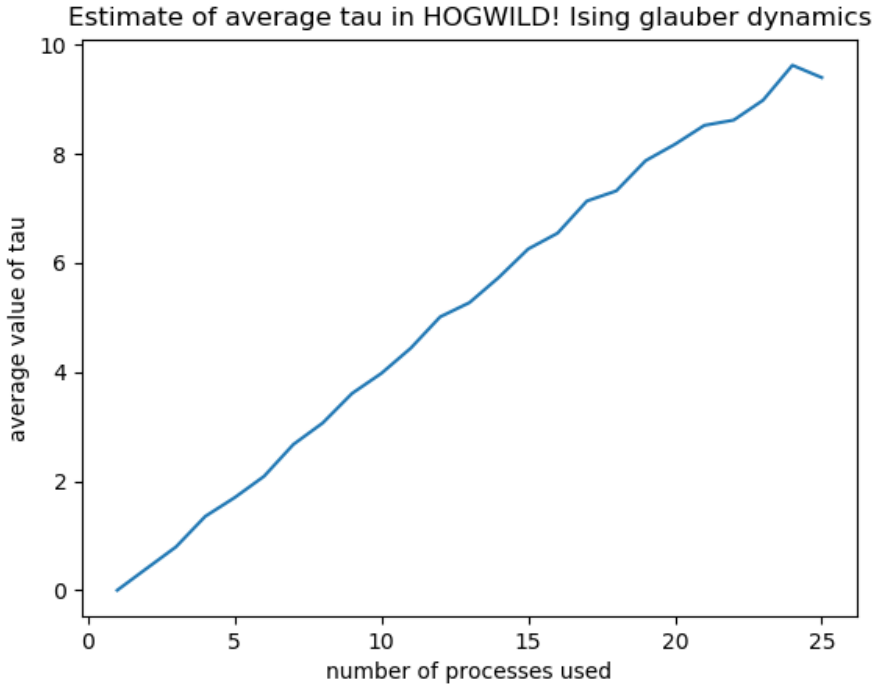

Figure 2: Average delay of reads for $CW(n, 0.5)$ model as the number of processors used varies.

complete bilinear function $f(X_1, \ldots, X_n) = \sum_{i \neq j} X_i X_j$ as we vary the number of nodes n in a Curie-Weiss model graph. Each red point is the empirical mean of the function f computed over 5000 samples from the HOGWILD! Markov chain corresponding to $CW(n, 0.5)$, and each blue point is the empirical mean produced from 5000 sequential runs of the same chain. Our theory (Theorem 5) predicts that the bias, the vertical difference in height between red and blue points, at any given value of n will be on the order of the standard deviation divided by $\sqrt{n}$ (standard deviation is $\Theta(n)$ and bias is $O(\sqrt{n})$). We plot error bars of this order, and find that the HOGWILD! means fall inside the error bars, thus corroborating our theory. We show that theory and practice coincide even for sparse graphs, by making the same plot for the $Grid(n, 0.5)$ model on the right of the same figure.

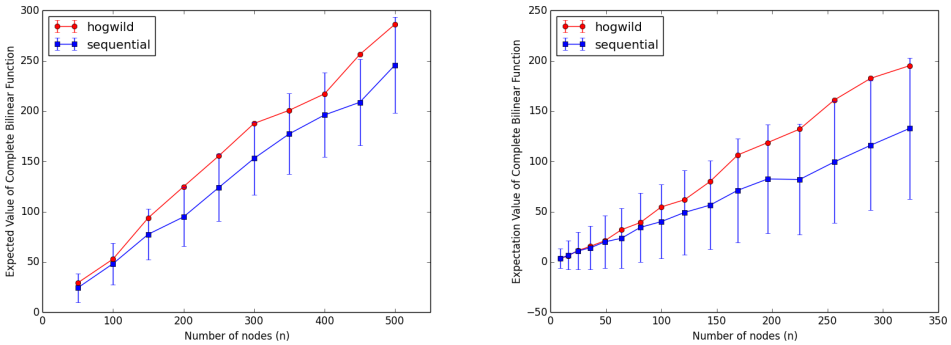

Figure 3: Means (with appropriately scaled error bars) of the complete bilinear function computed over 5000 sequential and hogwild runs of $CW(n, 0.5)$ (left) and $Grid(n, 0.5)$ (right).

6 Acknowledgements

We thank Prof. Srinivas Devdas and Xiangyao Yu for helping us gain access to and program on their multicore machines.

References

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

[DDK18] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

[DMR11] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1):149–189, 2011.

[DSOR16] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Ensuring rapid mixing and low bias for asynchronous gibbs sampling. In *JMLR workshop and conference proceedings*, volume 48, page 1567. NIH Public Access, 2016.

[DSZOR15] Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in neural information processing systems*, pages 2674–2682, 2015.

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[GG86] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. American Mathematical Society, 1986.

[GLP17] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121*, 2017.

[GSS18] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Higher order concentration for functions of weakly dependent random variables. *arXiv preprint arXiv:1801.06348*, 2018.

[JSW13] Matthew Johnson, James Saunderson, and Alan Willsky. Analyzing hogwild parallel gaussian gibbs sampling. In *Advances in Neural Information Processing Systems*, pages 2715–2723, 2013.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[LWR+15] Ji Liu, Stephen J Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *The Journal of Machine Learning Research*, 16(1):285–322, 2015.

[MBDC15] Ioannis Mitliagkas, Michael Borokhovich, Alexandros G Dimakis, and Constantine Caramanis. Frogwild!: fast pagerank approximations on graph engines. *Proceedings of the VLDB Endowment*, 8(8):874–885, 2015.

[MPP+15] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.

[MS10] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[NO14] Cyprien Noel and Simon Osindero. Dogwild!-distributed hogwild for cpu & gpu. In *NIPS Workshop on Distributed Machine Learning and Matrix Computations*, 2014.

[NRRW11] Feng Niu, Benjamin Recht, Christopher Re, and Stephen Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.

[RRWN11] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.

[SN10] Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

[TSD15] Alexander Terenin, Daniel Simpson, and David Draper. Asynchronous gibbs sampling. *arXiv preprint arXiv:1509.08999*, 2015.

[YHSD12] Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, and Inderjit Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 765–774. IEEE, 2012.

[ZR14] Ce Zhang and Christopher Ré. Dimmwitted: A study of main-memory statistical analytics. *Proceedings of the VLDB Endowment*, 7(12):1283–1294, 2014.