Ensuring Rapid Mixing and Low Bias for Asynchronous Gibbs Sampling

Christopher De Sa CDESA@STANFORD.EDU
Department of Electrical Engineering, Stanford University, Stanford, CA 94309

Kunle Olukotun KUNLE@STANFORD.EDU
Department of Electrical Engineering, Stanford University, Stanford, CA 94309

Christopher Ré CHRISMRE@STANFORD.EDU
Department of Computer Science, Stanford University, Stanford, CA 94309

Abstract

Gibbs sampling is a Markov chain Monte Carlo technique commonly used for estimating marginal distributions. To speed up Gibbs sampling, there has recently been interest in parallelizing it by executing asynchronously. While empirical results suggest that many models can be efficiently sampled asynchronously, traditional Markov chain analysis does not apply to the asynchronous case, and thus asynchronous Gibbs sampling is poorly understood. In this paper, we derive a better understanding of the two main challenges of asynchronous Gibbs: bias and mixing time. We show experimentally that our theoretical results match practical outcomes.

1. Introduction

Gibbs sampling is one of the most common Markov chain Monte Carlo methods used with graphical models (Koller & Friedman, 2009). In this setting, Gibbs sampling (Algorithm 1) operates iteratively by choosing at random a variable from the model at each timestep, and updating it by sampling from its conditional distribution given the other variables in the model. Often, it is applied to inference problems, in which we are trying to estimate the marginal probabilities of some query events in a given distribution.

For sparse graphical models, to which Gibbs sampling is often applied, each of these updates needs to read the values of only a small subset of the variables; therefore each update can be computed very quickly on modern hardware. Because of this and other useful properties of Gibbs sampling, many systems use Gibbs sampling to perform inference on big data (Newman et al., 2007; Lunn et al., 2009; McCallum et al., 2009; Smola & Narayanamurthy, 2010; Theis et al., 2012; Zhang & Ré, 2014).

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Algorithm 1 Gibbs sampling

Require: Variables x_i for $1 \leq i \leq n$, and distribution π.
  for $t = 1$ **to** T **do**
    Sample s uniformly from $\{1, \ldots, n\}$.
    Re-sample x_s uniformly from $\mathbf{P}_\pi(X_s | X_{\{1,\ldots,n\}\setminus\{s\}})$.
  end for

Since Gibbs sampling is such a ubiquitous algorithm, it is important to try to optimize its execution speed on modern hardware. Unfortunately, while modern computer hardware has been trending towards more parallel architectures (Sutter, 2005), traditional Gibbs sampling is an inherently sequential algorithm; that is, the loop in Algorithm 1 is not directly parallelizable. Furthermore, for sparse models, very little work happens within each iteration, meaning it is difficult to extract much parallelism from the body of this loop. Since traditional Gibbs sampling parallelizes so poorly, it is interesting to study variants of Gibbs sampling that can be parallelized. Several such variants have been proposed, including applications to latent Dirichlet allocation (Newman et al., 2007; Smola & Narayanamurthy, 2010) and distributed constraint optimization problems (Nguyen et al., 2013).

In one popular variant, multiple threads run the Gibbs sampling update rule in parallel without locks, a strategy called *asynchronous* or HOGWILD! execution—in this paper, we use these two terms interchangeably. This idea was proposed, but not analyzed theoretically, in Smola & Narayanamurthy (2010), and has been shown to give empirically better results on many models (Zhang & Ré, 2014). But when can we be sure that HOGWILD! Gibbs sampling will produce accurate results? Except for the case of Gaussian random variables (Johnson et al., 2013), there

is no existing analysis by which we can ensure that asynchronous Gibbs sampling will be appropriate for a particular application. Even the problems posed by HOGWILD!-Gibbs are poorly understood, and their solutions more so.

As we will show in the following sections, there are two main issues when analyzing asynchronous Gibbs sampling. Firstly, we will show by example that, surprisingly, HOGWILD!-Gibbs can be *biased*—unlike sequential Gibbs, it does not always produce samples that are arbitrarily close to the target distribution. Secondly, we will show that the *mixing time* (the time for the chain to become close to its stationary distribution) of asynchronous Gibbs sampling can be up to exponentially greater than that of the corresponding sequential chain.

To address the issue of bias, we need some way to describe the distance between the target distribution π and the distribution of the samples produced by HOGWILD!-Gibbs. The standard notion to use here is the *total variation distance*, but for the task of computing marginal probabilities, it gives an overestimate on the error caused by bias. To better describe the bias, we introduce a new notion of statistical distance, the *sparse variation distance*. While this relaxed notion of statistical distance is interesting in its own right, its main benefit here is that it uses a more local view of the chain to more tightly measure the effect of bias.

Our main goal is to identify conditions under which the bias and mixing time of asynchronous Gibbs can be bounded. One parameter that has been used to great effect in the analysis of Gibbs sampling is the *total influence* α of a model. The total influence measures the degree to which the marginal distribution of a variable can depend on the values of the other variables in the model—this parameter has appeared as part of a celebrated line of work on *Dobrushin's condition* ($\alpha < 1$), which ensures the rapid mixing of spin statistics systems (Dobrushin, 1956; Dyer et al., 2006; Hayes, 2006). It turns out that we can use this parameter to bound both the bias and mixing time of HOGWILD!-Gibbs, and so we make the following contributions:

- We describe a way to statistically model the asynchronicity in HOGWILD!-Gibbs sampling.
- To bound the bias, we prove that for classes of models with bounded total influence $\alpha = O(1)$, if sequential Gibbs sampling achieves small sparse variation distance to π in $O(n)$ steps, where n is the number of variables, then HOGWILD!-Gibbs samples achieve the same distance in at most $O(1)$ more steps.
- For models that satisfy Dobrushin's condition (that is, $\alpha < 1$), we show that the mixing time bounds of sequential and HOGWILD!-Gibbs sampling differ only by a factor of $1 + O(n^{-1})$.
- We validate our results experimentally and show that, by using asynchronous execution, we can achieve wall-clock speedups of up to $2.8\times$ on real problems.

2. Related Work

Much work has been done on the analysis of parallel Gibbs samplers. One simple way to parallelize Gibbs sampling is to run multiple chains independently in parallel: this heuristic uses parallelism to produce more samples overall, but does not produce accurate samples more quickly. Additionally, this strategy is sometimes worse than other strategies on a systems level (Smola & Narayanamurthy, 2010; Zhang & Ré, 2014), typically because it requires additional memory to maintain multiple models of the chain. Another strategy for parallelizing Gibbs sampling involves taking advantage of the structure of the underlying factor graph to run in parallel while still maintaining an execution pattern to which the standard sequential Gibbs sampling analysis can be applied (Gonzalez et al., 2011). Much further work has focused on parallelizing sampling for specific problems, such as LDA (Newman et al., 2007; Smola & Narayanamurthy, 2010) and others (Nguyen et al., 2013).

Our approach follows on the paper of Johnson et al. (2013), which named the HOGWILD!-Gibbs sampling algorithm and analyzed it for Gaussian models. Their main contribution is an analysis framework that includes a sufficient condition under which HOGWILD! Gaussian Gibbs samples are guaranteed to have the correct asymptotic mean. Recent work (Terenin et al., 2015) has analyzed a similar algorithm under even stronger regularity conditions. Here, we seek to give more general results for the analysis of HOGWILD!-Gibbs sampling on discrete-valued factor graphs.

The HOGWILD!-Gibbs sampling algorithm was inspired by a line of work on parallelizing stochastic gradient descent (SGD) by running it asynchronously. HOGWILD! SGD was first proposed by Niu et al. (2011), who proved that while running without locks causes race conditions, they do not significantly impede the convergence of the algorithm. The asynchronous execution strategy has been applied to many problems—such as PageRank approximations (Mitliagkas et al., 2015), deep learning (Noel & Osindero, 2014) and recommender systems (Yu et al., 2012)—so it is not surprising that it has been proposed for use with Gibbs sampling. Our goal in this paper is to combine analysis ideas that have been applied to Gibbs sampling and HOGWILD!, in order to characterize the behavior of asynchronous Gibbs. In particular, we are motivated by some recent work on the analysis of HOGWILD! for SGD (Liu et al., 2015; De Sa et al., 2015b; Mania et al., 2015; Liu & Wright, 2015). Several of these results suggest modeling the race conditions inherent in HOGWILD! SGD as noise in a stochastic process; this lets them bring a trove of statistical techniques to bear on the analysis of HOGWILD! SGD.

Therefore, in this paper, we will apply a similar stochastic process model to Gibbs sampling.

Several recent papers have focused on the mixing time of Gibbs sampling based on the structural properties of the model. Gotovos et al. (2015) and De Sa et al. (2015a) each show that Gibbs sampling mixes in polynomial time for a class of distributions bounded by some parameter. Unfortunately, these results both depend on *spectral methods* (that try to bound the spectral gap of the Markov transition matrix), which are difficult to apply to HOGWILD! Gibbs sampling for two reasons. First, spectral methods don't let us represent the sampler as a stochastic process, which limits the range of techniques we can use to model the noise. Secondly, while most spectral methods only apply to *reversible* Markov chains—and sequential Gibbs sampling is always a reversible chain—for HOGWILD!-Gibbs sampling the asynchronicity and parallelism make the chain non-reversible. Because of this, we were unable to use these spectral results in our asynchronous setting. We are forced to rely on the other method (Guruswami, 2000) for analyzing Markov processes, *coupling*—the type of analysis used with the Dobrushin condition—which we will describe in the following sections.

3. Modeling Asynchronicity

In this section, we describe a statistical model for asynchronous Gibbs sampling by adapting the hardware model outlined in De Sa et al. (2015b). Because we are motivated by the factor graph inference problem, we will focus on the case where the distribution π that we want to sample comes from a sparse, discrete graphical model.

Any HOGWILD!-Gibbs implementation involves some number of threads each repeatedly executing the Gibbs update rule on a single copy of the model (typically stored in RAM). We assume that this model serializes all writes, such that we can speak of the state of the system after t writes have occurred. We call this time t, and we will model the HOGWILD! system as a stochastic process adapted to the natural filtration $\mathcal{F}_t$. Here, $\mathcal{F}_t$ contains all events that have occurred up to time t, and we say an event is $\mathcal{F}_t$ *measurable* if it is known deterministically by time t.

We begin our construction by letting $x_{i,t}$ denote the ($\mathcal{F}_t$ measurable) value of variable i at time t, and letting $\tilde{I}_t$ be the ($\mathcal{F}_{t+1}$ measurable) index of the variable that we choose to sample at time t. For Gibbs sampling, we have

$$\forall i \in \{1, \ldots, n\},\ \mathbf{P}\left(\tilde{I}_t = i \middle| \mathcal{F}_t\right) = \frac{1}{n};$$

this represents the fact that we have an equal probability of sampling each variable.

Now that we have defined which variables are to be sampled, we proceed to describe how they are sampled. For HOGWILD!-Gibbs sampling, we must model the fact that the sampler does not get to use exactly the values of $x_{i,t}$; rather it has access to a cache containing potentially *stale* values. To do this, we define ($\mathcal{F}_{t+1}$ measurable) $\tilde{v}_{i,t} = x_{i,t-\tilde{\tau}_{i,t}}$, where $\tilde{\tau}_{i,t} \geq 0$ is a *delay parameter* ($\mathcal{F}_{t+1}$ measurable and independent of $\tilde{I}_t$) that represents how old the currently-cached value for variable i could be. A variable resampled using this stale data would have distribution

$$\mathbf{P}\left(\tilde{z}_{i,t} = z | \mathcal{F}_t\right) \propto \pi(\tilde{v}_{1,t}, \ldots, \tilde{v}_{i-1,t}, z, \tilde{v}_{i+1,t}, \ldots, \tilde{v}_{n,t}).$$

Using this, we can relate the values of the variables across time with

$$x_{i,t+1} = \begin{cases} \tilde{z}_{i,t} & \text{if } i = \tilde{I}_t \\ x_{i,t} & \text{otherwise.} \end{cases}$$

So far, our model is incompletely specified, because we have not described the distribution of the delays $\tilde{\tau}_{i,t}$. Unfortunately, since these delays depend on the number of threads and the specifics of the hardware (Niu et al., 2011), their distribution is difficult to measure. Instead of specifying a particular distribution, we require only a bound on the expected delay, $\mathbf{E}\left[\tilde{\tau}_{i,t} | \mathcal{F}_t\right] \leq \tau$. In this model, the τ parameter represents everything that is relevant about the hardware; representing the hardware in this way has been successful for the analysis of asynchronous SGD (Niu et al., 2011), so it is reasonable to use it for Gibbs sampling. In addition to this, we will need a similar parameter that bounds the tails of $\tilde{\tau}_{i,t}$ slightly more aggressively. We require that for some parameter τ^*, and for all i and t,

$$\mathbf{E}\left[\exp\left(n^{-1}\tilde{\tau}_{i,t}\right) \middle| \mathcal{F}_t\right] \leq 1 + n^{-1}\tau^*.$$

This parameter is typically very close to the expected value bound τ; in particular, as n approaches infinity, τ^* approaches τ.

4. The First Challenge: Bias

Perhaps the most basic result about sequential Gibbs sampling is the fact that, in the limit of large numbers of samples, it is unbiased. In order to measure convergence of Markov chains to their stationary distribution, it is standard to use the total variation distance.

Definition 1 (Total Variation Distance)**.** The *total variation distance* (Levin et al., 2009, p. 48) between two probability measures μ and ν on probability space Ω is defined as

$$\|\mu - \nu\|_{\text{TV}} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|,$$

that is, the maximum difference between the probabilities that μ and ν assign to a single event A.

It is a well-known result that, for Gibbs sampling on a strictly-positive target distribution π, it will hold that

$$\lim_{t\to\infty} \left\| P^{(t)}\mu_0 - \pi \right\|_{\mathrm{TV}} = 0, \tag{1}$$

where $P^{(t)}\mu_0$ denotes the distribution of the t-th sample.

One of the difficulties that arises when applying HOGWILD! to Gibbs sampling is that the race conditions from the asynchronous execution add bias to the samples — Equation 1 no longer holds. To understand why, we can consider a simple example.

4.1. Bias Example

Consider a simple model with two variables X_1 and X_2 each taking on values in $\{0, 1\}$, and having distribution

$$p(0,1) = p(1,0) = p(1,1) = \frac{1}{3} \qquad p(0,0) = 0.$$

Sequential Gibbs sampling on this model will produce unbiased samples from the target distribution. Unfortunately, this is not the case if we run HOGWILD!-Gibbs sampling on this model. Assume that the state is currently $(1, 1)$ and two threads, T_1 and T_2, simultaneously update X_1 and X_2 respectively. Since T_1 reads state $(1, 1)$ it will update X_1 to 0 or 1 each with probability 0.5; the same will be true for T_2 and X_2. Therefore, after this happens, every state will have probability 0.25; this includes the state $(0, 0)$ which should never occur! Over time, this race condition will produce samples with value $(0, 0)$ with some non-zero frequency; this is an example of *bias* introduced by the HOGWILD! sampling. Worse, this bias is not just theoretical: Figure 1 illustrates how the measured distribution for this model is affected by two-thread asynchronous execution. In particular, we observe that almost 5% of the mass is erroneously measured to be in the state $(0, 0)$, which has no mass at all in the true distribution. The total variation distance to the target distribution is quite large at 9.8%, and, unlike in the sequential case, this bias doesn't disappear as the number of samples goes to infinity.

4.2. Bounding the Bias

The previous example has shown that asynchronous Gibbs sampling will not necessarily produce a sequence of samples arbitrarily close to the target distribution. Instead, the samples may approach some other distribution, which we hope is sufficiently similar for some practical purpose. Often, the purpose of Gibbs sampling is to estimate the marginal distributions of individual variables or of events that each depend on only a small number of variables in the model. To characterize the accuracy of these estimates, the total variation distance is *too conservative*: it depends on the difference over all the events in the space, when most

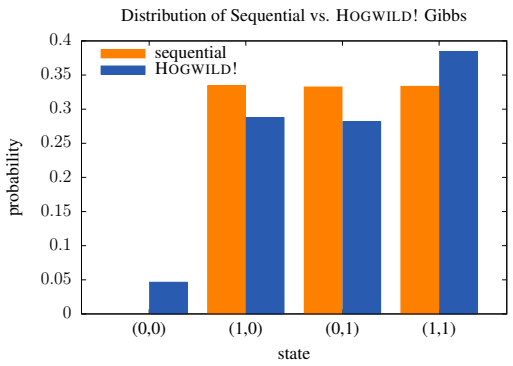

Figure 1. Bias introduced by HOGWILD!-Gibbs (10^6 samples).

of these are events that we do not care about. To address this, we introduce the following definition.

Definition 2 (Sparse Variation Distance)**.** For any event A in a probability space Ω over a set of variables V, let $|A|$ denote the number of variables upon which A depends. Then, for any two distributions μ and ν over Ω, we define the *ω-sparse variation distance* to be

$$\|\mu - \nu\|_{\mathrm{SV}(\omega)} = \max_{|A|\leq\omega} |\mu(A) - \nu(A)| .$$

For the wide variety of applications that use sampling for marginal estimation, the sparse variation distance measures the quantity we actually care about: the maximum possible bias in the marginal distribution of the samples. As we will show, asynchronous execution seems to have less effect on the sparse variation distance than the total variation distance, because sparse variation distance uses a more localized view of the chain. For example, in Figure 1, the total variation distance between the sequential and HOGWILD! distributions is 9.8%, while the 1-sparse variation distance is only 0.4%. That is, while HOGWILD! execution does introduce great bias into the distribution, it still estimates marginals of the individual variables accurately.

This definition suggests the question: how long do we have to run before our samples have low sparse variation distance from the target distribution? To answer this question, we introduce the following definition.

Definition 3 (Sparse Estimation Time)**.** The *ω-sparse estimation time* of a stochastic sampler with distribution $P^{(t)}\mu_0$ at time t and target distribution π is the first time t at which, for any initial distribution μ_0, the estimated distribution is within sparse variation distance ϵ of π,

$$t_{\mathrm{SE}(\omega)}(\epsilon) = \min\{t \in \mathbb{N} \mid \forall\mu_0,\ \|P^{(t)}\mu_0 - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon\}.$$

In many practical systems (Neubig, 2014; Shin et al., 2015), Gibbs sampling is used without a proof that it works;

instead, it is naively run for some fixed number of passes through the dataset. This naive strategy works for models for which accurate marginal estimates can be achieved after $O(n)$ samples. This $O(n)$ runtime is necessary for Gibbs sampling to be feasible on big data, meaning roughly that these are the models which it is interesting to try to speed up using asynchronous execution. Therefore, for the rest of this section, we will focus on the bias of the HOGWILD! chain for this class of models. When analyzing Gibbs sampling, we can bound the bias within the context of a coupling argument using a parameter called the *total influence*. While we arrived at this condition independently, it has been studied before, especially in the context of *Dobrushin's condition*, which ensures rapid mixing of Gibbs sampling.

Definition 4 (Total Influence)**.** Let π be a probability distribution over some set of variables I. Let B_j be the set of state pairs (X, Y) which differ only at variable j. Let $\pi_i(\cdot|X_{I\setminus\{i\}})$ denote the conditional distribution in π of variable i given all the other variables in state X. Then, define α, the total influence of π, as

$$\alpha = \max_{i\in I} \sum_{j\in I} \max_{(X,Y)\in B_j} \left\|\pi_i(\cdot|X_{I\setminus\{i\}}) - \pi_i(\cdot|Y_{I\setminus\{i\}})\right\|_{\mathrm{TV}}.$$

We say the model satisfies Dobrushin's condition if $\alpha < 1$.

One way to think of total influence for factor graphs is as a generalization of maximum degree; indeed, if a factor graph has maximum degree Δ, it can easily be shown that $\alpha \leq \Delta$. It turns out that if we can bound both this parameter and the sparse estimation time of sequential Gibbs sampling, we can give a simple bound on the sparse estimation time for asynchronous Gibbs sampling.

Claim 1. *Assume that we have a class of distributions with bounded total influence $\alpha = O(1)$. For each distribution π in the class, let $\bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon)$ be an upper bound on the ω-sparse estimation time of its sequential Gibbs sampler, and assume that it is a convex, decreasing function of ϵ. Further assume that, for any ϵ, across all models,*

$$\bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon) = O(n),$$

where n is the number of variables in the model. Then, for any ϵ, the sparse estimation time of HOGWILD!*-Gibbs across all models is bounded by*

$$t_{\mathrm{SE-hog}(\omega)}(\pi, \epsilon) \leq \bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon) + O(1).$$

Roughly, this means that HOGWILD!-Gibbs sampling "works" on all problems for which we know marginal estimation is "fast" and the total influence is bounded. Since the sparse estimation times here are measured in iterations, and the asynchronous sampler is able, due to parallelism, to run many more iterations in the same amount of wall clock time, this result implies that HOGWILD!-Gibbs can be much faster than sequential Gibbs for producing estimates of similar quality. To prove Claim 1, and more explicitly bound the bias, we use the following lemma.

Lemma 1. *Assume that we run* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Let $P^{(t)}_{\mathrm{hog}}$ denote the transition matrix of* HOGWILD!*-Gibbs and $P^{(t)}_{\mathrm{seq}}$ denote the transition matrix of sequential Gibbs. Then for any initial distribution μ_0 and for any t,*

$$\left\|P^{(t)}_{\mathrm{hog}}\mu_0 - P^{(t)}_{\mathrm{seq}}\mu_0\right\|_{\mathrm{SV}(\omega)} \leq \frac{\omega\alpha\tau t}{n^2}\exp\left(\frac{(\alpha-1)_+}{n}t\right),$$

where $(x)_+$ denotes x if $x > 0$ and 0 otherwise.

This lemma bounds the distance between the distributions of asynchronous and sequential Gibbs; if we let t be the sparse estimation time of sequential Gibbs, we can interpret this distance as an upper bound on the bias. When $t = O(n)$, this bias is $O(n^{-1})$, which has an intuitive explanation: for HOGWILD! execution, race conditions occur about once every $\Theta(n)$ iterations, so the bias is roughly proportional to the frequency of race conditions. This gives us a relationship between the statistical error of the algorithm and a more traditional notion of computational error.

Up until now, we have been assuming that we have a class for which the sparse estimation time is $O(n)$. Using the total influence α, we can identify a class of models known to meet this criterion.

Theorem 1. *For any distribution that satisfies Dobrushin's condition, $\alpha < 1$, the ω-sparse estimation time of the sequential Gibbs sampling process will be bounded by*

$$t_{\mathrm{SE-seq}(\omega)}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha}\log\left(\frac{\omega}{\epsilon}\right)\right\rceil.$$

This surprising result says that, in order to produce good marginal estimates for any model that satisfies Dobrushin's condition, we need only $O(n)$ samples! While we could now use Lemma 1 to bound the sparse estimation time for HOGWILD!-Gibbs, a more direct analysis produces a slightly better result, which we present here.

Theorem 2. *For any distribution that satisfies Dobrushin's condition, $\alpha < 1$, and for any ϵ that satisfies*

$$\epsilon \geq 2\omega\alpha\tau(1-\alpha)^{-1}n^{-1},$$

the ω-sparse estimation time of the HOGWILD! *Gibbs sampling process will be bounded by*

$$t_{\mathrm{SE-hog}(\omega)}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha}\log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)^2\epsilon}\right\rceil.$$

This result gives us a definite class of models for which HOGWILD!-Gibbs sampling is guaranteed to produce accurate marginal estimates quickly.

5. The Second Challenge: Mixing Times

Even though the HOGWILD!-Gibbs sampler produces biased estimates, it is still interesting to analyze how long we need to run it before the samples it produces are independent of its initial conditions. To measure the efficiency of a Markov chain, it is standard to use the *mixing time*.

Definition 5 (Mixing Time)**.** The *mixing time* (Levin et al., 2009, p. 55) of a stochastic process with transition matrix $P^{(t)}$ at time t and target distribution π is the first time t at which, for any initial distribution μ_0, the estimated distribution is within TV-distance ϵ of $P^{(t)}\pi$. That is,

$$t_{\text{mix}}(\epsilon) = \min\left\{t \middle| \forall \mu_0, \left\| P^{(t)}\mu_0 - P^{(t)}\pi \right\|_{\text{TV}} \leq \epsilon \right\}.$$

5.1. Mixing Time Example

As we did with bias, here we construct an example model for which asynchronous execution disastrously increases the mixing time. The model we will construct is rather extreme; we choose this model because simpler, practical models do not seem to exhibit this type of catastrophic increase in the mixing time. We start, for some odd constant N, with N variables $X_1, \ldots, X_N$ all in $\{-1, 1\}$, and one factor with energy

$$\phi_X(X) = -M_1 \left| \mathbf{1}^T X \right|,$$

for some very large energy parameter M_1. The resulting distribution will be almost uniform over all states with $\mathbf{1}^T X \in \{-1, 1\}$. To this model, we add another bank of variables $Y_1, \ldots, Y_N$ all in $\{-1, 1\}$. These variables also have a single associated factor with energy

$$\phi_Y(X, Y) = \begin{cases} \frac{\beta}{N}\left(\mathbf{1}^T Y\right)^2 & \text{if } \left|\mathbf{1}^T X\right| = 1 \\ M_2\left(\mathbf{1}^T Y\right)^2 & \text{if } \left|\mathbf{1}^T X\right| > 1 \end{cases},$$

for parameters β and M_2. Combining these two factors gives us the overall distribution for our model,

$$\pi(X, Y) = \frac{1}{Z}\exp\left(\phi_X(X) + \phi_Y(X, Y)\right),$$

where Z is the constant necessary for this to be a distribution. Roughly, the X dynamics are constructed to regularly "generate" race conditions, while the Y dynamics are chosen to "detect" these race conditions and mix very slowly as a result. This model is illustrated in Figure 2.

We simulated two-thread HOGWILD!-Gibbs on this model, measuring the marginal probability that $\mathbf{1}^T Y > 0$; by symmetry, this event has probability 0.5 in the stationary distribution for both the sequential and asynchronous samplers. Our results, for a model with $N = 2001$, $\beta = 0.3$, $M_1 = 10^{10}$, and $M_2 = 100$, and initial state $X = Y = \mathbf{1}$, are plotted in Figure 3. Notice that, while the sequential sampler achieves the correct marginal probability relatively quickly, the asynchronous samplers take a much longer time to achieve the correct result, even for a relatively small expected delay ($\tau = 0.5$). These results suggest that something catastrophic is happening to the mixing time when we switch from sequential to asynchronous execution — and in fact we can prove this is the case.

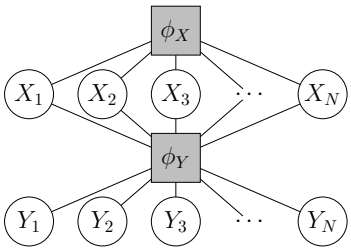

Figure 2. Factor graph model for mixing time example.

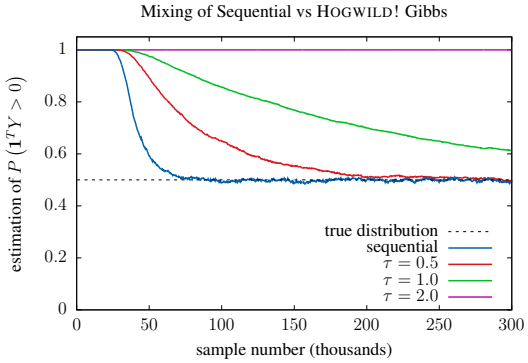

Figure 3. Example wherein asynchronous sampling greatly increases in mixing time. Marginals computed over 10^4 trials.

Statement 1. *For the example model described above, there exist parameters M_1, M_2, and β (as a function of N) such that the mixing time of sequential Gibbs sampling is $O(N \log N)$ but the mixing time of* HOGWILD!*-Gibbs sampling, even with $\tau = O(1)$, can be $\exp(\Omega(N))$.*

The intuition behind this statement is that for sequential Gibbs, the dynamics of the X part of the chain quickly causes it to have $\left|\mathbf{1}^T X\right| = 1$, and then remain there for the remainder of the simulation with high probability. This in turn causes the energy of the ϕ_Y factor to be essentially $\frac{\beta}{N}(\mathbf{1}^T Y)^2$, a model which is known to be fast-mixing because it satisfies Dobrushin's condition. On the other hand, for HOGWILD! Gibbs, due to race conditions we will see $\left|\mathbf{1}^T X\right| \neq 1$ with constant probability; this will cause the effective energy of the ϕ_Y factor to be dominated by the $M_2(\mathbf{1}^T Y)^2$ term, a model that is known to take exponential time to mix.

5.2. Bounding the Mixing Time

This example shows that fast mixing of the sequential sampler alone is not sufficient to guarantee fast mixing of the HOGWILD! chain. Consequently, we look for classes of models for which we can say something about the mixing time of both sequential and HOGWILD!-Gibbs. Dobrushin's condition is well known to imply rapid mixing of sequential Gibbs, and it turns out that we can leverage it again here to bound the mixing time of HOGWILD!-Gibbs.

Theorem 3. *Assume that we run Gibbs sampling on a distribution that satisfies Dobrushin's condition, $\alpha < 1$. Then the mixing time of sequential Gibbs will be bounded by*

$$t_{\mathrm{mix-seq}}(\epsilon) \leq \frac{n}{1-\alpha} \log\left(\frac{n}{\epsilon}\right).$$

Under the same conditions, the mixing time of HOGWILD!*-Gibbs will be bounded by*

$$t_{\mathrm{mix-hog}}(\epsilon) \leq \frac{n + \alpha\tau^*}{1-\alpha} \log\left(\frac{n}{\epsilon}\right).$$

The above example does not contradict this result since it does not satisfy Dobrushin's condition; in fact its total influence is very large and scales with n. We can compare these two mixing time results as

$$t_{\mathrm{mix-hog}}(\epsilon) \approx \left(1 + \alpha\tau^* n^{-1}\right) t_{\mathrm{mix-seq}}(\epsilon); \qquad (2)$$

the bounds on the mixing times differ by a negligible factor of $1 + O(n^{-1})$. This result shows that, for problems that satisfy Dobrusin's condition, HOGWILD!-Gibbs sampling mixes in about the same time as sequential Gibbs sampling, and is therefore a practical choice for generating samples.

5.3. A Positive Example: Ising Model

To gain intuition here, we consider a simple example. The Ising model (Ising, 1925) on a graph $G = (V, E)$ is a model over probability space $\{-1, 1\}^V$, and has distribution

$$p(\sigma) = \frac{1}{Z} \exp\left(\beta \sum_{(x,y)\in E} \sigma(x)\sigma(y) + \sum_{x\in V} B_x \sigma(x)\right),$$

where β is a parameter that is called the *inverse temperature*, the B_x are parameters that encode a *prior* on the variables, and Z is the normalization constant necessary for this to be a distribution. For graphs of maximum degree Δ and sufficiently small β, a bound on the mixing time of Gibbs sampling is known when $\Delta \tanh \beta \leq 1$. It turns out that the total influence of the Ising model can be bounded by $\alpha \leq \Delta \tanh \beta$, and so this condition is simply another way of writing Dobrushin's condition. We can therefore apply Theorem 3 to bound the mixing time of HOGWILD!-Gibbs with

$$t_{\mathrm{mix}}(\epsilon) \leq \frac{n + \tau^* \Delta \tanh \beta}{1 - \Delta \tanh \beta} \log\left(\frac{n}{\epsilon}\right).$$

This illustrates that the class of graphs we are considering includes some common, well-studied models.

5.4. Proof Outline

Here, we briefly describe the technique used to prove Theorem 3; for ease of presentation, we focus on the case where every variable takes on values in $\{-1, 1\}$. We start by introducing the idea of a coupling-based argument (Levin et al., 2009, p. 64), which starts by constructing two copies of the same Markov chain, X and $\bar{X}$, starting from different states but running together in the same probability space (i.e. using the same sources of randomness). For analyzing HOGWILD!-Gibbs sampling, we share randomness by having both chains sample the same variable at each iteration and sample it such that the resulting values are maximally correlated—additionally both chains are subject to the same HOGWILD! delays $\tilde{\tau}_{i,t}$.

At some random time, called the *coupling time* T_{c}, the chains will become equal—regardless of their initial conditions. Using this, we can bound the mixing time with

$$t_{\mathrm{mix}}(\epsilon) \leq \min\{t \mid \mathbf{P}(T_{\mathrm{c}} > t) \leq \epsilon\}.$$

In order to bound the probability that the chains are not equal at a particular time t, we focus on the quantity

$$\phi_t = \max_i P\left(X_{i,t} \neq \bar{X}_{i,t}\right). \qquad (3)$$

Under the conditions of Theorem 3, we are able to bound this using the total influence parameter. From here, we notice that by the union bound, $\mathbf{P}(T_{\mathrm{c}} > t) \leq n\phi_t$. Combining this with Equation 3 and reducing the subsequent expression lets us bound the mixing time, producing the result of Theorem 3.

6. Experiments

Now that we have derived a theoretical characterization of the behavior of HOGWILD!-Gibbs sampling, we examine whether this characterization holds up under experimental evaluation. First, we examine the mixing time claims we made in Section 5. Specifically, we want to check whether increasing the expected delay parameter τ^* actually increases the mixing time as predicted by Equation 2.

To do this, we simulated HOGWILD!-Gibbs sampling running on a random synthetic Ising model graph of order $n = 1000$, degree $\Delta = 3$, inverse temperature $\beta = 0.2$, and prior weights $E_x = 0$. This model has total influence $\alpha \leq 0.6$, and Theorem 3 guarantees that it will mix rapidly. Unfortunately, the mixing time of a chain is difficult to calculate experimentally. While techniques such as coupling from the past (Propp & Wilson, 1996) exist for estimating the mixing time, using these techniques in order to expose

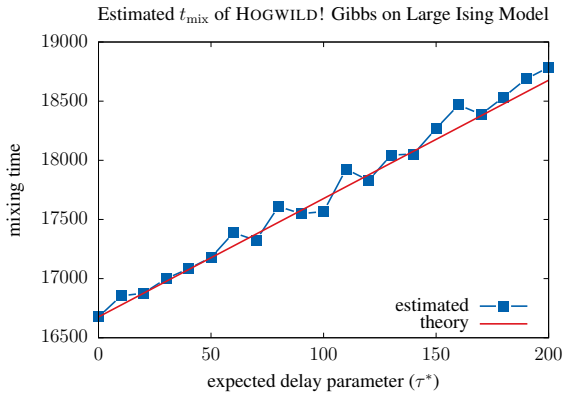

Figure 4. Comparison of estimated mixing time and theory-predicted (by Equation 2) mixing time as τ increases for a synthetic Ising model graph ($n = 1000$, $\Delta = 3$).

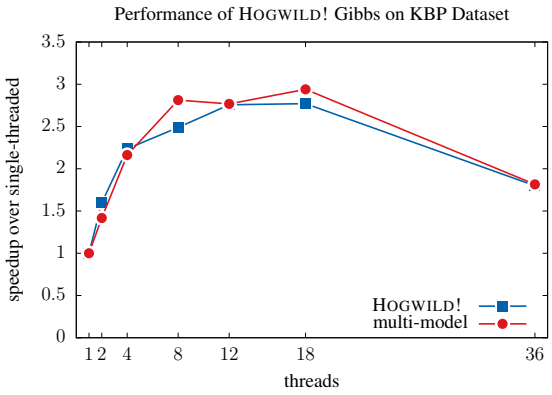

Figure 5. Speedup of HOGWILD! and multi-model Gibbs sampling on large KBP dataset (11 GB).

the (relatively small) dependence of the mixing time on τ proved to be computationally intractable.

Instead, we use a technique called coupling to the future. We initialize two chains, X and Y, by setting all the variables in X_0 to 1 and all the variables in Y_0 to -1. We proceed by simulating a coupling between the two chains, and return the coupling time T_{c}. Our estimate of the mixing time will then be $\hat{t}(\epsilon)$, where $\mathbf{P}(T_{\text{c}} \geq \hat{t}(\epsilon)) = \epsilon$.

Statement 2. *This experimental estimate is an upper bound for the mixing time. That is,* $\hat{t}(\epsilon) \geq t_{\text{mix}}(\epsilon)$.

To estimate $\hat{t}(\epsilon)$, we ran 10000 instances of the coupling experiment, and returned the sample estimate of $\hat{t}(1/4)$. To compare across a range of τ^*, we selected the $\tilde{\tau}_{i,t}$ to be independent and identically distributed according to the maximum-entropy distribution supported on $\{0, 1, \ldots, 200\}$ consistent with a particular assignment of τ^*. The resulting estimates are plotted as the blue series in Figure 4. The red line represents the mixing time that would be predicted by naively applying Equation 2 using the estimate of the sequential mixing time as a starting point — we can see that it is a very good match for the experimental results. This experiment shows that, at least for one archetypal model, our theory accurately characterizes the behavior of HOGWILD! Gibbs sampling as the delay parameter τ^* is changed, and that using HOGWILD!-Gibbs doesn't cause the model to catastrophically fail to mix.

Of course, in order for HOGWILD!-Gibbs to be useful, it must also speed up the execution of Gibbs sampling on some practical models. It is already known that this is the case, as these types of algorithms been widely implemented in practice (Smyth et al., 2009; Smola & Narayanamurthy, 2010). To further test this, we ran HOGWILD!-Gibbs sampling on a real-world 11 GB Knowledge Base Population dataset (derived from the TAC-KBP challenge) using a machine with a single-socket, 18-core Xeon E7-8890 CPU and 1 TB RAM. As a comparison, we also ran a "multi-model" Gibbs sampler: this consists of multiple threads with a single execution of Gibbs sampling running independently in each thread. This sampler will produce the same number of samples as HOGWILD!-Gibbs, but will require more memory to store multiple copies of the model.

Figure 5 reports the speedup, in terms of wall-clock time, achieved by HOGWILD!-Gibbs on this dataset. On this machine, we get speedups of up to $2.8\times$, although the program becomes memory-bandwidth bound at around 8 threads, and we see no significant speedup beyond this. With any number of workers, the run time of HOGWILD!-Gibbs is close to that of multi-model Gibbs, which illustrates that the additional cache contention caused by the HOGWILD! updates has little effect on the algorithm's performance.

7. Conclusion

We analyzed HOGWILD!-Gibbs sampling, a heuristic for parallelized MCMC sampling, on discrete-valued graphical models. First, we constructed a statistical model for HOGWILD!-Gibbs by adapting a model already used for the analysis of asynchronous SGD. Next, we illustrated a major issue with HOGWILD!-Gibbs sampling: that it produces biased samples. To address this, we proved that if for some class of models with bounded total influence, only $O(n)$ sequential Gibbs samples are necessary to produce good marginal estimates, then HOGWILD!-Gibbs sampling produces equally good estimates after only $O(1)$ additional steps. Additionally, for models that satisfy Dobrushin's condition ($\alpha < 1$), we proved mixing time bounds for sequential and asynchronous Gibbs sampling that differ by only a factor of $1 + O(n^{-1})$. Finally, we showed that our theory matches experimental results, and that HOGWILD!-Gibbs produces speedups up to $2.8\times$ on a real dataset.

Acknowledgments

The authors acknowledge the support of: DARPA FA8750-12-2-0335; NSF IIS-1247701; NSF CCF-1111943; DOE 108845; NSF CCF-1337375; DARPA FA8750-13-2-0039; NSF IIS-1353606; ONR N000141210041 and N000141310129; NIH U54EB020405; Oracle; NVIDIA; Huawei; SAP Labs; Sloan Research Fellowship; Moore Foundation; American Family Insurance; Google; and Toshiba.

"The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, NSF, ONR, NIH, or the U.S. Government."

References

De Sa, Christopher, Zhang, Ce, Olukotun, Kunle, and Ré, Christopher. Rapidly mixing gibbs sampling for a class of factor graphs using hierarchy width. In *NIPS*. NIPS Foundation, 2015a.

De Sa, Christopher, Zhang, Ce, Olukotun, Kunle, and Ré, Christopher. Taming the wild: A unified analysis of HOGWILD!-style algorithms. In *NIPS*. NIPS Foundation, 2015b.

Dobrushin, RL. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(4):329–383, 1956.

Dyer, Martin, Goldberg, Leslie Ann, and Jerrum, Mark. Dobrushin conditions and systematic scan. In *in Proc. 10th International Workshop on Randomization and Computation, Lecture Notes in Computer Science 4110*, pp. 327–338. Springer, 2006.

Gonzalez, Joseph, Low, Yucheng, Gretton, Arthur, and Guestrin, Carlos. Parallel gibbs sampling: From colored fields to thin junction trees. In *AISTATS*, pp. 324–332, 2011.

Gotovos, Alkis, Hassani, Hamed, and Krause, Andreas. Sampling from probabilistic submodular models. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1936–1944. Curran Associates, Inc., 2015.

Guruswami, Venkatesan. Rapidly mixing markov chains: A comparison of techniques. *Available: cs. washington. edu/homes/venkat/pubs/papers. html*, 2000.

Hayes, Thomas P. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 39–46. IEEE, 2006.

Ising, Ernst. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

Johnson, Matthew, Saunderson, James, and Willsky, Alan. Analyzing hogwild parallel gaussian gibbs sampling. In *NIPS*, pp. 2715–2723, 2013.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Levin, David Asher, Peres, Yuval, and Wilmer, Elizabeth Lee. *Markov chains and mixing times*. American Mathematical Soc., 2009.

Liu, Ji and Wright, Stephen J. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIOPT*, 25(1):351–376, 2015.

Liu, Ji, Wright, Stephen J, Ré, Christopher, Bittorf, Victor, and Sridhar, Srikrishna. An asynchronous parallel stochastic coordinate descent algorithm. *JMLR*, 16:285–322, 2015.

Lunn, David, Spiegelhalter, David, Thomas, Andrew, and Best, Nicky. The BUGS project: evolution, critique and future directions. *Statistics in medicine*, (25):3049–3067, 2009.

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

McCallum, Andrew, Schultz, Karl, and Singh, Sameer. Factorie: Probabilistic programming via imperatively defined factor graphs. In *NIPS*, pp. 1249–1257, 2009.

Mitliagkas, Ioannis, Borokhovich, Michael, Dimakis, Alexandros G., and Caramanis, Constantine. Frogwild!: Fast pagerank approximations on graph engines. *PVLDB*, 2015.

Neubig, Graham. Simple, correct parallelization for blocked gibbs sampling. Technical report, Nara Institute of Science and Technology, 2014.

Newman, David, Smyth, Padhraic, Welling, Max, and Asuncion, Arthur U. Distributed inference for latent dirichlet allocation. In *NIPS*, pp. 1081–1088, 2007.

Nguyen, Duc Thien, Yeoh, William, and Lau, Hoong Chuin. Distributed gibbs: A memory-bounded sampling-based dcop algorithm. In *Proceedings of the 2013 international conference on Autonomous agents*

and multi-agent systems, pp. 167–174. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

Niu, Feng, Recht, Benjamin, Re, Christopher, and Wright, Stephen. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pp. 693–701, 2011.

Noel, Cyprien and Osindero, Simon. Dogwild!–Distributed Hogwild for CPU & GPU. 2014.

Propp, James Gary and Wilson, David Bruce. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252, 1996.

Shin, Jaeho, Wu, Sen, Wang, Feiran, De Sa, Christopher, Zhang, Ce, Wang, Feiran, and Ré, Christopher. Incremental knowledge base construction using deepdive. *PVLDB*, 2015.

Smola, Alexander and Narayanamurthy, Shravan. An architecture for parallel topic models. *PVLDB*, 2010.

Smyth, Padhraic, Welling, Max, and Asuncion, Arthur U. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems*, pp. 81–88, 2009.

Sutter, Herb. The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. *Dr. Dobb's Journal*, 30(3), 2005.

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

Theis, Lucas, Sohl-dickstein, Jascha, and Bethge, Matthias. Training sparse natural image models with a fast gibbs sampler of an extended state space. In *NIPS*, pp. 1124–1132. 2012.

Yu, Hsiang-Fu, Hsieh, Cho-Jui, Si, Si, and Dhillon, Inderjit S. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *ICDM*, pp. 765–774, 2012.

Zhang, Ce and Ré, Christopher. DimmWitted: A study of main-memory statistical analytics. *PVLDB*, 2014.

A. Additional Bias Results

In this section, we present the following additional result that bounds the sparse estimation time of general Gibbs samplers. In particular, this theorem provides an explicit form of the result given in Claim 1.

Theorem 4. *Assume that we run* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Let $\bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon)$ be some upper bound on the ω-sparse estimation time of the corresponding sequential chain, and assume that it is a convex and decreasing function of ϵ. For any $\epsilon > 0$, define*

$$c = \frac{1}{n}\bar{t}_{\mathrm{SE-seq}(\omega)}\left(\frac{\epsilon}{2}\right).$$

Then, as long as ϵ is large enough that

$$\epsilon \geq \frac{2\omega\alpha\tau c}{n}e^{c\cdot(\alpha-1)_+},$$

where we use the notation $(x)_+ = \max(0, x)$, the ω-sparse estimation time of the HOGWILD! *chain can be bounded with*

$$t_{\mathrm{SE-hog}(\omega)}(\epsilon) \leq \left\lceil \bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon) + \frac{2\omega\alpha\tau c^2}{\epsilon}e^{c\cdot(\alpha-1)_+} \right\rceil.$$

B. Proofs

Here, we provide proofs for the results in the paper. In the first subsection, we will state lemmas and known results that we will use in the subsequent proofs. Next, we will prove the Claims and Theorems stated in the body of the paper. Finally, we will prove the lemmas previously stated.

B.1. Statements of Lemmas

First, we state a proposition from Levin et al. (2009). This proposition relates the concept of a coupling with the total variation distance between the distributions of two random variables.

Proposition 1 (Proposition 4.7 from Levin et al. (2009))**.** *Let X and Y be two random variables that take on values in the same set, and let their distributions be μ and ν, respectively. Then for any coupling, $(\bar{X}, \bar{Y})$ it will hold that*

$$\|\mu - \nu\|_{\mathrm{TV}} \leq P\left(\bar{X} \neq \bar{Y}\right).$$

Furthermore, there exists a coupling for which equality is achieved; this is called an optimal *coupling.*

We can prove a related result for sparse variation distance.

Lemma 2. *Let X and Y be two random variables that each assign values to a set of variables $\{1, \ldots, n\}$, and let their distributions be μ and ν, respectively. Then for any coupling, $(\bar{X}, \bar{Y})$ it will hold that*

$$\|\mu - \nu\|_{\mathrm{SV}(\omega)} \leq \max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right).$$

We state a lemma that bounds the expected total variation distance between the marginal distributions of two states using the total influence α. Note that a similar statement to that proved in this lemma may be used as an alternate definition for the total influence α; the definition given in the body of the paper is used because it is more intuitive and does not require introducing the concept of a coupling. This lemma will be useful later when proving the subsequent lemmas stated in this subsection.

Lemma 3. *If π is a distribution with total influence α, and X and Y are two random variables that take on values in the state space of π, then for any variable i*

$$\mathbf{E}\left[\|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}}\right] \leq \alpha \max_j P\left(X_j \neq Y_j\right),$$

where, for simplicity of notation, we let $\pi_i(\cdot|X)$ denote the conditional distribution of variable i in π given the values of all the other variables in state X.

Next, we state three lemmas, each of which give bounds on the quantity

$$P\left(X_{i,t} \neq Y_{i,t}\right)$$

for some coupling of two (potentially asynchronous) Gibbs sampling chains. First, we state the result for comparing two synchronous chains.

Lemma 4. *Consider sequential Gibbs sampling on a distribution π with total influence α. Then, for any initial states (X_0, Y_0) there exists a coupling of the chains (X_t, Y_t) such that for any variable i and any time t,*

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right).$$

Second, we state the result comparing two HOGWILD! chains.

Lemma 5. *Consider any model of* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Then, for any initial states (X_0, Y_0) there exists a coupling (X_t, Y_t) of the* HOGWILD!*-Gibbs sampling chains starting at X_0 and Y_0 respectively such that for any variable i and any time t,*

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

Third, we state the result comparing a sequential and an asynchronous chain.

Lemma 6. *Consider any model of* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Then if for any initial states (X_0, Y_0) we can construct a coupling (X_t, Y_t) such that the process X_t is distributed according to the dynamics of* HOGWILD!*-Gibbs, the process Y_t is distributed according to the dynamics of sequential Gibbs, and for any time t,*

$$\max_i P\left(X_{i,t+1} \neq Y_{i,t+1}\right) \leq \left(1 - \frac{1-\alpha}{n}\right) \max_i P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha\tau}{n^2}.$$

As a secondary result, if the chain satisfies Dobrushin's condition ($\alpha < 1$), then for any variable i and any time t,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\alpha\tau}{(1-\alpha)n}.$$

Lemma 7 (Monotonic Sequence Domination Lemma)**.** *Let $x_0, x_1, \ldots$ be a sequence such that, for all t,*

$$x_{t+1} \leq f_t(x_t, x_{t-1}, \ldots, x_0),$$

where f_t is a function that is monotonically increasing in all of its arguments. Then, for any sequence $y_0, y_1, \ldots$, if $x_0 = y_0$ and for all t,

$$y_{t+1} \geq f_t(y_t, y_{t-1}, \ldots, y_0),$$

then for all t,

$$x_t \leq y_t.$$

Lemma 8. *Consider the model on N variables X_i, for N odd, where each X_i takes on values in $\{-1, 1\}$ and has probability*

$$\pi(X) = \frac{1}{Z_X} \begin{cases} 1 & \text{if } \left|\mathbf{1}^T X\right| = 1 \\ 0 & \text{if } \left|\mathbf{1}^T X\right| > 1 \end{cases}$$

Then Gibbs sampling on this model (assuming that we allow the chain to start only at a state X where $\pi(X) > 0$) has mixing time

$$t_{\text{mix}} = O(n \log n).$$

B.2. Proofs of Bias Results

First, we restate and prove Claim 1. This proof will use the result of Theorem 4, which we will prove subsequently. We note here that the use of a convex upper bound for the sparse estimation time of the sequential chain (as opposed to using the sequential chain's sparse estimation time directly) is an unfortunate consequence of the proof—we hope that a more careful analysis could remove it or replace it with a more natural condition.

Claim 1. *Assume that we have a class of distributions with bounded total influence $\alpha = O(1)$. For each distribution π in the class, let $\bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon)$ be an upper bound on the ω-sparse estimation time of its sequential Gibbs sampler, and assume that it is a convex, decreasing function of ϵ. Further assume that, for any ϵ, across all models,*

$$\bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon) = O(n),$$

where n is the number of variables in the model. Then, for any ϵ, the sparse estimation time of HOGWILD!*-Gibbs across all models is bounded by*

$$t_{\mathrm{SE-hog}(\omega)}(\pi, \epsilon) \leq \bar{t}_{\mathrm{SE-seq}(\omega)}(\pi, \epsilon) + O(1).$$

Proof. First, note that, since $\alpha = O(1)$, we know by the definition of big-O notation that for some α^*, for all models in the class, the total influence of that model will be $\alpha \leq \alpha^*$. Similarly, since we assumed that, for any ϵ and across all models π,

$$\bar{t}_{\mathrm{SM-seq}(\omega)}(\pi, \epsilon) = O(n),$$

then for each ϵ, there must exist a $c(\epsilon)$ such that for any distribution π with n variables in the class,

$$t_{\mathrm{SM-seq}(\omega)}(\pi, \epsilon) \leq n \cdot c(\epsilon).$$

For some error ϵ and model π, we would like to apply Theorem 4 to bound its mixing time. In order to apply the theorem, we must satisfy the conditions on ϵ: it suffices for

$$n \geq \frac{2\omega\alpha^*\tau c(\epsilon/2)}{\epsilon} \exp\left((\alpha^* - 1)_+ c(\epsilon/2)\right).$$

Under this condition, applying the theorem allows us to bound the ω-sparse estimation time of the HOGWILD! chain with

$$\begin{aligned} t_{\mathrm{SE-hog}(\omega)}(\epsilon) &\leq \left\lceil \bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon) + \frac{2\omega\alpha^*\tau c(\epsilon/2)^2}{\epsilon} \exp\left((\alpha^* - 1)_+ c(\epsilon/2)\right) \right\rceil \\ &\leq \bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon) + \frac{2\omega\alpha^*\tau c(\epsilon/2)^2}{\epsilon} \exp\left((\alpha^* - 1)_+ c(\epsilon/2)\right) + 1 \end{aligned}$$

Therefore, if we define

$$N(\epsilon) = \frac{2\omega\alpha^*\tau c(\epsilon/2)}{\epsilon} \exp\left((\alpha^* - 1)_+ c(\epsilon/2)\right),$$

and

$$T(\epsilon) = \frac{2\omega\alpha^*\tau c(\epsilon/2)^2}{\epsilon} \exp\left((\alpha^* - 1)_+ c(\epsilon/2)\right) + 1,$$

then it follows that, for any ϵ and for all models with $n \geq N(\epsilon)$,

$$t_{\mathrm{SM-hog}(\omega)}(\epsilon) \leq t_{\mathrm{SM-seq}(\omega)}(\epsilon) + T(\epsilon).$$

This is equivalent to saying that, for any ϵ and across all models,

$$t_{\mathrm{SM-hog}(\omega)}(\epsilon) \leq t_{\mathrm{SM-seq}(\omega)}(\epsilon) + O(1).$$

This proves the claim. □

Next, we restate and prove the bias lemma, Lemma 1.

Lemma 1. *Assume that we run* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Let $P_{\text{hog}}^{(t)}$ denote the transition matrix of* HOGWILD!*-Gibbs and $P_{\text{seq}}^{(t)}$ denote the transition matrix of sequential Gibbs. Then for any initial distribution μ_0 and for any t,*

$$\left\| P_{\text{hog}}^{(t)}\mu_0 - P_{\text{seq}}^{(t)}\mu_0 \right\|_{\text{SV}(\omega)} \leq \frac{\omega\alpha\tau t}{n^2} \exp\left(\frac{(\alpha-1)_+}{n}t\right),$$

where $(x)_+$ denotes x if $x > 0$ and 0 otherwise.

Proof of Lemma 1. We start by using the primary result from Lemma 6. This result states that we can construct a coupling (X_t, Y_t) of the HOGWILD! and sequential chains starting at any initial distributions X_0 and Y_0 such that at any time t,

$$\max_i P\left(X_{i,t+1} \neq Y_{i,t+1}\right) \leq \left(1 - \frac{1-\alpha}{n}\right) \max_i P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha\tau}{n^2}.$$

Now, for any initial distribution μ_0, assume that we start with $X_0 = Y_0$, where both are distributed according to μ_0. Then, trivially,

$$P\left(X_{i,0} \neq Y_{i,0}\right) = 0.$$

It follows from recursive application of the sub-result of Lemma 6 that, for this coupling,

$$\begin{aligned}
\max_i P\left(X_{i,t} \neq Y_{i,t}\right) &\leq \sum_{k=0}^{t-1} \left(1 + \frac{\alpha-1}{n}\right)^k \frac{\alpha\tau}{n^2} \\
&\leq t\left(1 + \frac{(\alpha-1)_+}{n}\right)^t \frac{\alpha\tau}{n^2} \\
&\leq \exp\left(\frac{(\alpha-1)_+}{n}t\right)\frac{\alpha\tau t}{n^2},
\end{aligned}$$

where $(x)_+$ denotes $\max(0, x)$. It follows by the union bound that, for any set of variables I with $|I| \leq \omega$, the probability that the coupling is unequal in at least one of those variables is

$$\begin{aligned}
P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) &\leq \omega \max_i P\left(X_{i,t} \neq Y_{i,t}\right) \\
&\leq \exp\left(\frac{(\alpha-1)_+}{n}t\right)\frac{\omega\alpha\tau t}{n^2}.
\end{aligned}$$

Since this inequality holds for any set of variable I with $|I| \leq \omega$, it follows that

$$\begin{aligned}
\max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) &\leq \omega \max_i P\left(X_{i,t} \neq Y_{i,t}\right) \\
&\leq \exp\left(\frac{(\alpha-1)_+}{n}t\right)\frac{\omega\alpha\tau t}{n^2}.
\end{aligned}$$

We can proceed to apply Lemma 2, which lets us conclude that

$$\begin{aligned}
\left\| P_{\text{hog}}^{(t)}\mu_t - P_{\text{seq}}^{(t)}\nu_t \right\|_{\text{SV}(\omega)} &\leq \omega \max_i P\left(X_{i,t} \neq Y_{i,t}\right) \\
&\leq \exp\left(\frac{(\alpha-1)_+}{n}t\right)\frac{\omega\alpha\tau t}{n^2}
\end{aligned}$$

This is the desired result. □

Next, we restate and prove the full bias result, Theorem 4.

Theorem 4. *Assume that we run* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Let $\bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon)$ be some upper bound on the ω-sparse estimation time of the corresponding sequential chain, and assume that it is a convex and decreasing function of ϵ. For any $\epsilon > 0$, define*

$$c = \frac{1}{n}\bar{t}_{\mathrm{SE-seq}(\omega)}\left(\frac{\epsilon}{2}\right).$$

Then, as long as ϵ is large enough that

$$\epsilon \geq \frac{2\omega\alpha\tau c}{n}e^{c\cdot(\alpha-1)_+},$$

where we use the notation $(x)_+ = \max(0, x)$, the ω-sparse estimation time of the HOGWILD! *chain can be bounded with*

$$t_{\mathrm{SE-hog}(\omega)}(\epsilon) \leq \left\lceil \bar{t}_{\mathrm{SE-seq}(\omega)}(\epsilon) + \frac{2\omega\alpha\tau c^2}{\epsilon}e^{c\cdot(\alpha-1)_+} \right\rceil.$$

Proof of Theorem 4. We start with the result of Lemma 1, which lets us conclude that

$$\|\mu_t - \nu_t\|_{\mathrm{SV}(\omega)} \leq \exp\left(\frac{(\alpha-1)_+}{n}t\right)\frac{\omega\alpha\tau t}{n^2}$$

where $\mu_t = P_{\mathrm{hog}}^{(t)}\mu_0$ and $\nu_t = P_{\mathrm{seq}}^{(t)}\mu_0$ are the distributions of the HOGWILD! and sequential Gibbs sampling chains, respectively, starting in state μ_0. Next, since ν_t has the dynamics of the sequential Gibbs sampling chain, and $\bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon)$ is an upper bound for the sparse estimation time, it follows that for any ϵ, if

$$t \geq \bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon),$$

then

$$\|\nu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since $\bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon)$ is a decreasing function of ϵ, it must have an inverse function. Furthermore, since it is convex, its inverse function must also be convex. Therefore, we can also write the above expression in terms of the inverse function; for any t,

$$\|\nu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \bar{t}^{-1}_{\mathrm{SM-seq}(\omega)}(t).$$

Therefore, by the triangle inequality, for any t,

$$\begin{aligned}\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} &\leq \|\mu_t - \nu_t\|_{\mathrm{SV}(\omega)} + \|\nu_t - \pi\|_{\mathrm{SV}(\omega)} \\ &\leq \frac{\omega\alpha\tau t}{n^2}\exp\left(\frac{(\alpha-1)_+}{n}t\right) + \bar{t}^{-1}_{\mathrm{SM-seq}(\omega)}(t).\end{aligned}$$

Now, for any particular ϵ, let

$$t_0 = \bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon),$$

and let

$$t_1 = \bar{t}_{\mathrm{SM-seq}(\omega)}\left(\frac{\epsilon}{2}\right).$$

Further define

$$R = \frac{\omega\alpha\tau t_1}{n^2}\exp\left(\frac{(\alpha-1)_+}{n}t_1\right).$$

Therefore, for any $t_0 \leq t \leq t_1$,

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq R + \bar{t}^{-1}_{\mathrm{SM-seq}(\omega)}(t).$$

By convexity of $\bar{t}^{-1}_{\mathrm{SM-seq}(\omega)}$, we can bound this expression over the interval $t_0 \leq t \leq t_1$ with

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq R + \frac{t_1 - t}{t_1 - t_0}\cdot\epsilon + \frac{t - t_0}{t_1 - t_0}\cdot\frac{\epsilon}{2},$$

and so, if we want this to be less than ϵ, it suffices to choose t such that

$$\epsilon = R + \frac{t_1 - t}{t_1 - t_0} \cdot \epsilon + \frac{t - t_0}{t_1 - t_0} \cdot \frac{\epsilon}{2}$$

which will occur when

$$t = t_0 + \frac{2R(t_1 - t_0)}{\epsilon}.$$

Now, applying the definition

$$c = \frac{1}{n} \bar{t}_{\mathrm{SM-seq}(\omega)} \left(\frac{\epsilon}{2} \right) = \frac{t_1}{n}$$

lets us equivalently write R as

$$R = \frac{\omega \alpha \tau c}{n} \exp\left(\frac{(\alpha - 1)_+}{n} t_1 \right).$$

Recall that as a condition for the theorem, we assumed that

$$\epsilon \geq \frac{2\omega \alpha \tau c}{n} \exp\left(\frac{(\alpha - 1)_+}{n} t_1 \right).$$

It follows from this and our expression for R that

$$R \leq \frac{\epsilon}{2}.$$

Therefore this assignment of t will satisfy the previous constraint that $t_0 \leq t \leq t_1$, and so for this assignment of t, and for any initial distribution μ_0, it holds that

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Therefore, by the definition of sparse estimation time, the sparse estimation time of the HOGWILD! chain will be

$$t_{\mathsf{SE-hog}(\omega)}(\epsilon) \leq t,$$

for this assignment of t. Now, recall that above we assigned

$$t = t_0 + \frac{2R(t_1 - t_0)}{\epsilon}.$$

Under this condition, we can bound this whole error term as

$$\frac{2R(t_1 - t_0)}{\epsilon} \leq \frac{2\omega \alpha \tau t_1^2}{n^2 \epsilon} \exp\left(\frac{(\alpha - 1)_+}{n} t_1 \right).$$

Combining this with the definitions of t_0 and c lets us state that

$$t \leq \bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon) + \frac{2\omega \alpha \tau c^2}{\epsilon} \exp\left((\alpha - 1)_+ c \right).$$

Taking the ceiling implies that, when

$$t = \left\lceil \bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon) + \frac{2\omega \alpha \tau c^2}{\epsilon} \exp\left((\alpha - 1)_+ c \right) \right\rceil,$$

for any initial distribution μ_0,

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since we above defined μ_t to be the distribution of HOGWILD! Gibbs after t timesteps, $\mu_t = P^{(t)} \mu_0$, where $P^{(t)}$ is the transition matrix of HOGWILD! Gibbs after t timesteps. We can thus equivalently write this as

$$\left\| P^{(t)} \mu_0 - \pi \right\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Therefore, by the definition of sparse estimation time,

$$t_{\mathsf{SE-hog}}(\epsilon) \leq \left\lceil \bar{t}_{\mathrm{SM-seq}(\omega)}(\epsilon) + \frac{2\omega \alpha \tau c^2}{\epsilon} \exp\left((\alpha - 1)_+ c \right) \right\rceil.$$

This proves the theorem. □

Next, we restate and prove the theorem that bounds the sparse estimation time of sequential Gibbs for distributions that satisfy Dobrushin's condition.

Theorem 1. *For any distribution that satisfies Dobrushin's condition, $\alpha < 1$, the ω-sparse estimation time of the sequential Gibbs sampling process will be bounded by*

$$t_{\mathrm{SE-seq}(\omega)}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha} \log\left(\frac{\omega}{\epsilon}\right) \right\rceil.$$

Proof of Theorem 1. We start by using the result of Lemma 4. This result states that, for any initial distributions (X_0, Y_0), there exists a coupling (X_t, Y_t) of the sequential Gibbs sampling chains starting at distributions X_0 and Y_0, respectively, such that for any variable i and any time t,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right).$$

It follows by the union bound that, for any set of variables I with $|I| \leq \omega$, the probability that the coupling is unequal in at least one of those variables is

$$P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right).$$

Since this inequality holds for any set of variable I with $|I| \leq \omega$, it follows that

$$\max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right).$$

We can proceed to apply Lemma 2, which lets us conclude that, if we let μ_t and ν_t denote the distributions of X_t and Y_t, respectively, then

$$\|\mu_t - \nu_t\|_{\mathrm{SV}(\omega)} \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right).$$

Since this was true for any initial distributions for X_0 and Y_0, it will hold in particular for Y_0 distributed according to π, the stationary distribution of the chain. In this case, $\nu_t = \pi$, and so for any initial distribution μ_0 for X_0,

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right).$$

Now, in order for this to be bounded by ϵ, it suffices to choose t such that

$$\omega \exp\left(-\frac{1-\alpha}{n}t\right) \leq \epsilon.$$

This will occur whenever

$$t \geq \frac{n}{1-\alpha} \log\left(\frac{\omega}{\epsilon}\right)$$

(here we used the fact that $\alpha < 1$ to do the division). Taking the ceiling, we can conclude that when

$$t = \left\lceil \frac{n}{1-\alpha} \log\left(\frac{\omega}{\epsilon}\right) \right\rceil.$$

for any initial distribution μ_0,

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since we defined μ_t to be the distribution of X_t, it must hold that $\mu_t = P^{(t)}\mu_0$, where μ_0 is the initial distribution of X_0, and $P^{(t)}$ is the transition matrix associated with running t steps of sequential Gibbs sampling. Thus, we can rewrite this as

$$\left\| P^{(t)}\mu_0 - \pi \right\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since this result held for any initial assignment of X_0 and therefore for any μ_0, by the definition of sparse estimation time it follows that

$$t_{\mathsf{SE-seq}}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha} \log\left(\frac{\omega}{\epsilon}\right) \right\rceil.$$

This proves the theorem. □

Next, we restate and prove the theorem that bounds the sparse estimation time of HOGWILD! Gibbs for distributions that satisfy Dobrushin's condition.

Theorem 2. *For any distribution that satisfies Dobrushin's condition, $\alpha < 1$, and for any ϵ that satisfies*

$$\epsilon \geq 2\omega\alpha\tau(1-\alpha)^{-1}n^{-1},$$

the ω-sparse estimation time of the HOGWILD! *Gibbs sampling process will be bounded by*

$$t_{\mathrm{SE-hog}(\omega)}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha} \log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)^2\epsilon} \right\rceil.$$

Proof of Theorem 2. We start by using the secondary result from Lemma 6—we can safely use this result because we assumed the chain satisfied Dobrushin's condition ($\alpha < 1$). This result states that we can construct a coupling (X_t, Y_t) of the HOGWILD! and sequential chains starting at any initial distributions X_0 and Y_0 such that at any time t,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\alpha\tau}{(1-\alpha)n}.$$

It follows by the union bound that, for any set of variables I with $|I| \leq \omega$, the probability that the coupling is unequal in at least one of those variables is

$$P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\omega\alpha\tau}{(1-\alpha)n}$$

Since this inequality holds for any set of variable I with $|I| \leq \omega$, it follows that

$$\max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ X_{i,t} \neq Y_{i,t}\right) \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\omega\alpha\tau}{(1-\alpha)n}.$$

We can proceed to apply Lemma 2, which lets us conclude that, if we let μ_t and ν_t denote the distributions of X_t and Y_t respectively,

$$\|\mu_t - \nu_t\|_{\mathrm{SV}(\omega)} \leq \omega \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\omega\alpha\tau}{(1-\alpha)n}.$$

To bound the sparse estimation time, notice that for any fixed ϵ (independent of n), in order to achieve

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon,$$

it suffices to choose any t such that

$$\omega \exp\left(-\frac{1-\alpha}{n}t\right) \leq \epsilon - \frac{\omega\alpha\tau}{(1-\alpha)n}.$$

This will occur when

$$\frac{1-\alpha}{n}t \geq \log\left(\frac{\omega}{\epsilon}\right) - \log\left(1 - \frac{\omega\alpha\tau}{(1-\alpha)n\epsilon}\right).$$

Next, recall that we assumed that

$$\epsilon \geq \frac{2\omega\alpha\tau}{(1-\alpha)n};$$

therefore ϵ is large enough that

$$\frac{\omega\alpha\tau}{(1-\alpha)n\epsilon} \leq \frac{1}{2}.$$

It is easy to prove that, for all $x \leq \frac{1}{2}$,

$$\log(1-x) \geq 2x.$$

Therefore, under this condition in ϵ, it suffices to choose t such that

$$\frac{1-\alpha}{n}t \geq \log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)n\epsilon};$$

this will occur whenever

$$t \geq \frac{n}{1-\alpha}\log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)^2\epsilon}.$$

Taking the ceiling implies that, when

$$t = \left\lceil \frac{n}{1-\alpha}\log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)^2\epsilon} \right\rceil,$$

for any initial distribution μ_0,

$$\|\mu_t - \pi\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since we defined μ_t above to be the distribution of X_t, it follows that $\mu_t = P^{(t)}\mu_0$, where μ_0 is the initial distribution of X_0 and $P^{(t)}$ is the transition matrix associated with running t steps of HOGWILD! Gibbs. Therefore, we can rewrite this as

$$\left\| P^{(t)}\mu_0 - \pi \right\|_{\mathrm{SV}(\omega)} \leq \epsilon.$$

Since this is true for any initial distribution of X_0 and therefore for any μ_0, it follows from the definition of sparse estimation time that

$$t_{\mathsf{SE-hog}}(\epsilon) \leq \left\lceil \frac{n}{1-\alpha}\log\left(\frac{\omega}{\epsilon}\right) + \frac{2\omega\alpha\tau}{(1-\alpha)^2\epsilon} \right\rceil.$$

This proves the theorem. □

B.3. Proofs of Mixing Time Results

First, we restate and prove Statement 1.

Statement 1. *For the example model described above, there exist parameters M_1, M_2, and β (as a function of N) such that the mixing time of sequential Gibbs sampling is $O(N \log N)$ but the mixing time of* HOGWILD!*-Gibbs sampling, even with $\tau = O(1)$, can be $\exp(\Omega(N))$.*

Proof of Statement 1. We start out by proving that the model mixes rapidly in the sequential case.

First, we assume that we select M_1 large enough that, even for potentially exponential run times, the dynamics of the chain are indistinguishable from the chain with $M_1 = \infty$. In particular, this alternate chain will have the following properties:

- The dynamics of the X part of the chain do not depend in any way on the value of Y.
- If at any point, $\left|\mathbf{1}^T X\right| > 1$, whenever we sample an X variable, we will re-sample it if possible to decrease the value of $\left|\mathbf{1}^T X\right|$ with probability 1.
- As long as $\left|\mathbf{1}^T X\right| = 1$ at some point in time, this will remain true, and the dynamics of the X part of the chain will be those of the chain described in Lemma 8.

We assume that we choose M_1 large enough that these properties hold over all time windows discussed in this proof with high probability.

Now, by the coupon collector's problem, after $O(N \log N)$ timesteps, we have sampled all the variables with high probability. If we have sampled all the variables with high probability, then we will certainly have $\left|\mathbf{1}^T X\right| = 1$ with high probability.

Once we have $\left|\mathbf{1}^T X\right| = 1$, Lemma 8 ensures that, after $O(N \log N)$ additional timesteps, the X part of the chain will be close to its stationary distribution.

Meanwhile, while $\left|\mathbf{1}^T X\right| = 1$, the dynamics of the Y part of the chain are exactly Gibbs sampling over the model with energy

$$\phi_Y(Y) = \frac{\beta}{N}\left(\mathbf{1}^T Y\right)^2.$$

For any $\beta < 1$, this is known to mix in $O(N \log N)$ time, since it satisfies Dobrushin's condition. Therefore, after $O(N \log N)$ steps after we have $\left|\mathbf{1}^T X\right| = 1$, the Y part of the chain will also be close to its stationary distribution.

Summing up the times for the above events gives us a total mixing time for this chain of

$$t_{\mathrm{mix-seq}} = O(N \log N).$$

Next we prove that the model takes a potentially exponential time to mix in the asynchronous case. Assume here that our model of execution has two threads, which always either sample two X variables independently and asynchronously, or sample a single Y variable synchronously (i.e. there is never any delay when reading the value of a Y variable). For this execution pattern, we have uniformly that $\tau_{i,t} \leq 1$. In particular, this has $\tau = O(1)$.

Now, consider the case where the two threads each choose to sample a variable in X that can be switched. Since at least $\frac{1}{4}$ of the variables are variables in X that can be switched, this will occur with probability at least $\frac{1}{16}$. Given this, they will each independently switch their variable with probability $\frac{1}{2}$. This means that both variables are switched with probability $\frac{1}{4}$ — but this would place the system in a state where

$$\left|\mathbf{1}^T X\right| > 1.$$

At any time when $\left|\mathbf{1}^T X\right| = 1$, this will occur with probability $\frac{1}{64}$, which implies that whenever we sample Y, the probability that $\left|\mathbf{1}^T X\right| > 1$ is at least $\frac{1}{64}$.

Now, assume without loss of generality that we initialize Y such that $\mathbf{1}^T Y = N$. Let ρ_t denote the value of $\mathbf{1}^T Y$ at time t. Assuming that we sample a variable Y_i with value 1, while $\left|\mathbf{1}^T X\right| = 1$, the probability that it will be switched will be

$$\begin{aligned} P\left(\text{value switched}\right) &= \frac{\exp\left(\beta n^{-1}(\rho_t - 1)^2\right)}{\exp\left(\beta n^{-1}(\rho_t - 1)^2\right) + \exp\left(\beta n^{-1}(\rho_t)^2\right)} \\ &= \left(1 + \exp\left(\beta n^{-1}\left((\rho_t)^2 - (\rho_t - 1)^2\right)\right)\right)^{-1} \\ &= \left(1 + \exp\left(\beta n^{-1}(2\rho_t - 1)\right)\right)^{-1}. \end{aligned}$$

Note that since $\rho_t \leq N$ at all times, if $\beta < 1$,

$$\beta N^{-1}(2\rho_t - 1) \leq 2.$$

We also can verify that, for any $0 \leq x \leq 2$, as a basic property of the exponential function,

$$\left(1 + \exp(x)\right)^{-1} \leq \frac{1}{2} - \frac{x}{6}.$$

Therefore, as long as $\rho_t > 0$,

$$P\left(\text{value switched}\right) \leq \frac{1}{2} - \frac{\beta \rho_t}{3n}.$$

Therefore, as long as $\rho_t > 0$, and $\left|\mathbf{1}^T X\right| = 1$,

$$\begin{aligned} \mathbf{E}\left[\rho_{t+1} | \mathcal{F}_t\right] &\geq \rho_t + 2\left(\frac{N - \rho_t}{2N} - \frac{1}{2} + \frac{\beta \rho_t}{3N}\right) \\ &= \rho_t + 2\left(\frac{-\rho_t}{2N} + \frac{\beta \rho_t}{3N}\right) \\ &= \rho_t \left(1 - \frac{3 - 2\beta}{3N}\right). \end{aligned}$$

On the other hand, if $\left|\mathbf{1}^T X\right| > 1$, then we can pick M_2 large enough such that with high probability, as long as $\rho_t > 0$, all variables Y_i are always sampled to be 1. In this case,

$$\begin{aligned}\mathbf{E}\left[\rho_{t+1}|\mathcal{F}_t\right] &\geq \rho_t + 2\left(\frac{N-\rho_t}{2N}\right) \\ &= \rho_t\left(1-\frac{1}{N}\right)+1.\end{aligned}$$

In general, since $\left|\mathbf{1}^T X\right| > 1$ with probability at least $\frac{1}{64}$,

$$\begin{aligned}\mathbf{E}\left[\rho_{t+1}|\mathcal{F}_t\right] &\geq \left(1-\frac{1}{64}\right)\rho_t\left(1-\frac{3-2\beta}{3N}\right)+\frac{1}{64}\left(\rho_t\left(1-\frac{1}{N}\right)+1\right) \\ &= \rho_t\left(1-\left(1-\frac{1}{64}\right)\frac{3-2\beta}{3N}-\frac{1}{64N}\right)+\frac{1}{64} \\ &= \rho_t\left(1-\frac{1}{N}+\left(1-\frac{1}{64}\right)\frac{2\beta}{3N}\right)+\frac{1}{64} \\ &\geq \rho_t\left(1-\frac{1}{N}\right)+\frac{1}{64}\end{aligned}$$

This expression has fixed point

$$\rho^* = \frac{N}{64}.$$

Since ρ is written as a sum of independent samples, as long as $\rho > 0$, the distribution of ρ is going to be exponentially concentrated around its expected value, which we have just shown is at least $\frac{N}{64}$. It follows that it is exponentially unlikely to ever achieve a value of ρ that is not positive. By the union bound, there is some $t = \exp(\Omega(N))$ such that, after t timesteps, $\rho_t > 0$ with high probability.

But, the actual probability that $\rho > 0$ in the stationary distribution is exactly $\frac{1}{2}$, by symmetry. It follows that the mixing time for the HOGWILD! chain must be greater than t; that is,

$$t_{\mathrm{mix-hog}} \geq \exp(\Omega(N)).$$

This finishes our proof of the statement. □

Next, we restate and prove Theorem 3.

Theorem 3. *Assume that we run Gibbs sampling on a distribution that satisfies Dobrushin's condition, $\alpha < 1$. Then the mixing time of sequential Gibbs will be bounded by*

$$t_{\mathrm{mix-seq}}(\epsilon) \leq \frac{n}{1-\alpha}\log\left(\frac{n}{\epsilon}\right).$$

Under the same conditions, the mixing time of HOGWILD!*-Gibbs will be bounded by*

$$t_{\mathrm{mix-hog}}(\epsilon) \leq \frac{n+\alpha\tau^*}{1-\alpha}\log\left(\frac{n}{\epsilon}\right).$$

Proof of First Part of Theorem 3. If we use the coupling from Lemma 4, then by the result of that lemma,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right),$$

It follows by the union bound that

$$P\left(X_t \neq Y_t\right) \leq n\exp\left(-\frac{1-\alpha}{n}t\right).$$

Now, assume that we initialize X_0 with distribution μ_0, and Y_0 with the stationary distribution π. By Proposition 1, since X_t has distribution $P^{(t)}\mu_0$ and Y_t has distribution $P^{(t)}\pi$, this is equivalent to saying

$$\left\|P^{(t)}\mu_0 - P^{(t)}\pi\right\|_{\text{TV}} \leq n \exp\left(-\frac{1-\alpha}{n}t\right).$$

Therefore, in order for

$$\left\|P^{(t)}\mu_0 - P^{(t)}\pi\right\|_{\text{TV}} \leq \epsilon,$$

it suffices to choose t such that

$$\epsilon = n \exp\left(-\frac{1-\alpha}{n}t\right).$$

This occurs when

$$t = \frac{n}{1-\alpha}\log\left(\frac{n}{\epsilon}\right),$$

which is the desired expression. □

Proof of Second Part of Theorem 3. If we use the coupling from Lemma 5, then by the result of that lemma,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right),$$

It follows by the union bound that

$$P\left(X_t \neq Y_t\right) \leq n \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

Next, recall that we assumed that our HOGWILD!-Gibbs sampler has target distribution π. Now, assume that we initialize X_0 with distribution μ_0, and Y_0 with the target distribution π. By Proposition 1, since X_t has distribution $P^{(t)}\mu_0$ and Y_t has distribution $P^{(t)}\pi$, this is equivalent to saying

$$\left\|P^{(t)}\mu_0 - P^{(t)}\pi\right\|_{\text{TV}} \leq n \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

Therefore, in order for

$$\left\|P^{(t)}\mu_0 - P^{(t)}\pi\right\|_{\text{TV}} \leq \epsilon,$$

it suffices to choose t such that

$$\epsilon = n \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

This occurs when

$$t = \frac{n+\alpha\tau^*}{1-\alpha}\log\left(\frac{n}{\epsilon}\right),$$

which is the desired expression. □

Next, we restate and prove Statement 2, which says that our experimental strategy provides a valid upper bound on the mixing time.

Statement 2. *This experimental estimate is an upper bound for the mixing time. That is,* $\hat{t}(\epsilon) \geq t_{\text{mix}}(\epsilon)$.

Proof of Statement 2. Consider the partial ordering of states in this Ising model defined by

$$Y \preceq X \leftrightarrows \forall i,\, Y_i \leq X_i.$$

Next, consider the coupling procedure that, at each time t, chooses a random variable $\tilde{I}_t$ to sample and a random $\tilde{R}_t$ uniformly on $[0,1]$. It then computes p_t, the marginal probability of sampling the chosen variable as 1, and assigns the variable as

$$\text{new value of } X_{\tilde{I}_t} = \begin{cases} 1 & \text{if } \tilde{R}_t < p_t, \\ 0 & \text{otherwise} \end{cases}.$$

This sampling procedure is equivalent to the one that we use in the experiment, and it will produce a chain that is consistent with the Ising model's dynamics.

If we consider the evolution of two coupled chains $X^{(t)}$ and $Y^{(t)}$ using the same values of $\tilde{I}_t$ and $\tilde{R}_t$, then from the way that we constructed the coupling, it follows that if

$$Y^{(0)} \preceq X^{(0)},$$

then for any future time t,

$$Y^{(t)} \preceq X^{(t)}.$$

This is because if

$$Y^{(t)} \preceq X^{(t)},$$

then the marginal probability of assigning 1 to any particular variable in X is always no less than the marginal probability of assigning 1 to the same variable in Y.

Therefore, if we initialize all $X_i^{(0)} = 1$ and all $Y_i^{(0)} = -1$, and run the coupling until time T_{coupling}, the time at which

$$Y^{(T_{\text{coupling}})} = X^{(T_{\text{coupling}})},$$

then by the previous analysis, since for any chain U initialized at any state $U^{(0)}$,

$$Y^{(0)} \preceq U^{(0)} \preceq X^{(0)},$$

it follows that

$$Y^{(T_{\text{coupling}})} \preceq U^{(T_{\text{coupling}})} \preceq X^{(T_{\text{coupling}})},$$

and so,

$$Y^{(T_{\text{coupling}})} = U^{(T_{\text{coupling}})} = X^{(T_{\text{coupling}})}.$$

Since this was true for any initial value of U, it follows that T_{coupling} is a coupling time for any two initial values of the chain. Therefore, by Corollary 5.3 from Levin et al. (2009),

$$\max_{\mu_0} \left\| P^{(t)}\mu_0 - \pi \right\|_{\text{TV}} \leq P\left(T_{\text{coupling}} > t\right).$$

If we use our definition of $\hat{t}(\epsilon)$ where

$$P\left(T_{\text{coupling}} > \hat{t}(\epsilon)\right) = \epsilon,$$

then this implies that

$$\max_{\mu_0} \left\| P^{(\hat{t})}\mu_0 - \pi \right\|_{\text{TV}} \leq \epsilon.$$

This in turn implies that $\hat{t}$ is a upper bound on the mixing time, which is the desired result. □

B.4. Proofs of Lemmas

In this section, we will restate and prove the lemmas used earlier in the appendix.

Lemma 2. *Let X and Y be two random variables that each assign values to a set of variables $\{1, \ldots, n\}$, and let their distributions be μ and ν, respectively. Then for any coupling, $(\bar{X}, \bar{Y})$ it will hold that*

$$\|\mu - \nu\|_{\text{SV}(\omega)} \leq \max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right).$$

Proof of Lemma 2. For any set of variables $I \subset \{1, \ldots, n\}$, let $M_I(\mu)$ denote the marginal distribution of the variables in I in the distribution μ. In particular, M_I includes all events A that depend only on variables in set I. Next, let $\bar{X}_I$ and $\bar{Y}_I$ denote the values of $\bar{X}$ and $\bar{Y}$ on those variables in I; this will be a coupling of the distributions $M_I(\mu)$ and $M_I(\nu)$. Therefore, by Proposition 1,

$$\|M_I(\mu) - M_I(\nu)\|_{\text{SV}(\omega)} \leq P\left(\bar{X}_I \neq \bar{Y}_I\right) = P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right).$$

Let Ω_I denote all events in the original probability space Ω that depend only on the variables in I. By the definition of total variation distance,

$$\|M_I(\mu) - M_I(\nu)\|_{\mathrm{SV}(\omega)} = \max_{A \in \Omega_I} |\mu(A) - \nu(A)| .$$

Therefore,

$$\max_{A \in \Omega_I} |\mu(A) - \nu(A)| \leq P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right).$$

Now, since this was true for any I, it is certainly true if we maximize both sides over all I with $|I| \leq \omega$. Therefore,

$$\max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} \max_{A \in \Omega_I} |\mu(A) - \nu(A)| \leq \max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right).$$

The left side can be reduced to

$$\max_{|A| \leq \omega} |\mu(A) - \nu(A)| \leq \max_{I \subseteq \{1,\ldots,n\},\, |I| \leq \omega} P\left(\exists i \in I,\ \bar{X}_i \neq \bar{Y}_i\right)$$

and applying the definition of sparse variation distance proves the lemma. □

Lemma 3. *If π is a distribution with total influence α, and X and Y are two random variables that take on values in the state space of π, then for any variable i*

$$\mathbf{E}\left[\|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}}\right] \leq \alpha \max_j P\left(X_j \neq Y_j\right),$$

where, for simplicity of notation, we let $\pi_i(\cdot|X)$ denote the conditional distribution of variable i in π given the values of all the other variables in state X.

Proof of Lemma 3. Let n be the number of variables in the model. For all $k \in \{0, 1, \ldots, n\}$, let Z_k be a random variable that takes on values in the state space of π such that, for all $j \in \{1, \ldots, n\}$,

$$Z_{k,j} = \begin{cases} X_j & \text{if } j > k \\ Y_j & \text{if } j \leq k \end{cases}.$$

In particular, $Z_0 = X$ and $Z_n = Y$. Now, by the triangle inequality on the total variation distance,

$$\begin{aligned} \|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}} &= \|\pi_i(\cdot|Z_0) - \pi_i(\cdot|Z_n)\|_{\mathrm{TV}} \\ &\leq \sum_{k=1}^{n} \|\pi_i(\cdot|Z_{k-1}) - \pi_i(\cdot|Z_k)\|_{\mathrm{TV}} \end{aligned}$$

Next, we note that $Z_{k-1} = Z_k$ if and only if $X_k = Y_k$. Therefore,

$$\|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}} \leq \sum_{k=1}^{n} \mathbf{1}_{X_k \neq Y_k} \|\pi_i(\cdot|Z_{k-1}) - \pi_i(\cdot|Z_k)\|_{\mathrm{TV}} .$$

Since Z_{k-1} and Z_k differ only at most at index k, it follows that $(Z_{k-1}, Z_k) \in B_k$, and so,

$$\|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}} \leq \sum_{k=1}^{n} \mathbf{1}_{X_k \neq Y_k} \max_{(U,V) \in B_k} \|\pi_i(\cdot|U) - \pi_i(\cdot|V)\|_{\mathrm{TV}} .$$

Maximizing over the right side produces

$$\|\pi_i(\cdot|X) - \pi_i(\cdot|Y)\|_{\mathrm{TV}} \leq \max_j \sum_{k=1}^{n} \mathbf{1}_{X_k \neq Y_k} \max_{(U,V) \in B_k} \|\pi_j(\cdot|U) - \pi_j(\cdot|V)\|_{\mathrm{TV}} .$$

Taking the expected value of both sides produces

$$\begin{aligned}\mathbf{E}\left[\|\pi_i(\cdot|X)-\pi_i(\cdot|Y)\|_{\mathrm{TV}}\right] &\le \max_j \sum_{k=1}^{n} \mathbf{E}\left[\mathbf{1}_{X_k\neq Y_k}\right] \max_{(U,V)\in B_k} \|\pi_j(\cdot|U)-\pi_j(\cdot|V)\|_{\mathrm{TV}} \\ &= \max_j \sum_{k=1}^{n} P\left(X_k\neq Y_k\right) \max_{(U,V)\in B_k} \|\pi_j(\cdot|U)-\pi_j(\cdot|V)\|_{\mathrm{TV}} \\ &\le \left(\max_k P\left(X_k\neq Y_k\right)\right)\left(\max_j \sum_{k=1}^{n} \max_{(U,V)\in B_k} \|\pi_j(\cdot|U)-\pi_j(\cdot|V)\|_{\mathrm{TV}}\right).\end{aligned}$$

Finally, applying the definition of total influence gives us

$$\mathbf{E}\left[\|\pi_i(\cdot|X)-\pi_i(\cdot|Y)\|_{\mathrm{TV}}\right] \le \alpha \max_k P\left(X_k\neq Y_k\right).$$

This proves the lemma. □

Lemma 4. *Consider sequential Gibbs sampling on a distribution π with total influence α. Then, for any initial states (X_0, Y_0) there exists a coupling of the chains (X_t, Y_t) such that for any variable i and any time t,*

$$P\left(X_{i,t}\neq Y_{i,t}\right) \le \exp\left(-\frac{1-\alpha}{n}t\right).$$

Proof of Lemma 4. Define the coupling as follows. Start in state (X_0, Y_0), and at each timestep, choose a single variable i uniformly at random for both chains to sample. Then, sample the selected variable in both chains using the optimal coupling, of the conditional distributions of the variable to be sampled in both chains, guaranteed by Proposition 1. Iterated over time, this defines a full coupling of the two chains.

Next, consider the event that $X_{i,t+1}\neq Y_{i,t+1}$. This event will occur if one of two things happens: either we didn't sample variable i at time t and $X_{i,t}\neq Y_{i,t}$; or we did sample variable i at time t, and the sampled variables were not equal. Since the probability of sampling variable i is $\frac{1}{n}$, and we know the probability that the sampled variables were not equal from Proposition 1, it follows that, by the law of total probability,

$$P\left(X_{i,t+1}\neq Y_{i,t+1}\right) = \left(1-\frac{1}{n}\right)P\left(X_{i,t}\neq Y_{i,t}\right) + \frac{1}{n}\mathbf{E}\left[\|\pi_i(\cdot|X_t)-\pi_i(\cdot|Y_t)\|_{\mathrm{TV}}\right],$$

where $\pi_i(\cdot|X_t)$ denotes the conditional distribution of variable i in π given the values of the other variables in X_t.

Next, we apply the Lemma 3, which gives us

$$P\left(X_{i,t+1}\neq Y_{i,t+1}\right) \le \left(1-\frac{1}{n}\right)P\left(X_{i,t}\neq Y_{i,t}\right) + \frac{\alpha}{n}\max_j P\left(X_{j,t}\neq Y_{j,t}\right).$$

Maximizing both sides over i produces

$$\begin{aligned}\max_i P\left(X_{i,t+1}\neq Y_{i,t+1}\right) &\le \left(1-\frac{1}{n}\right)\max_i P\left(X_{i,t}\neq Y_{i,t}\right) + \frac{\alpha}{n}\max_j P\left(X_{j,t}\neq Y_{j,t}\right) \\ &= \left(1-\frac{1}{n}+\frac{\alpha}{n}\right)\max_i P\left(X_{i,t}\neq Y_{i,t}\right).\end{aligned}$$

Applying this inequality recursively, and noting that $\max_i P\left(X_{i,0}\neq Y_{i,0}\right)\le 1$, we get

$$\max_i P\left(X_{i,t}\neq Y_{i,t}\right) \le \left(1-\frac{1-\alpha}{n}\right)^t \le \exp\left(-\frac{1-\alpha}{n}t\right).$$

This gives us the desired result. □

Lemma 5. *Consider any model of* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Then, for any initial states (X_0, Y_0) there exists a coupling (X_t, Y_t) of the* HOGWILD!*-Gibbs sampling chains starting at X_0 and Y_0 respectively such that for any variable i and any time t,*

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

Proof of Lemma 5. Define the coupling as follows. Start in state (X_0, Y_0), and at each timestep, choose a single variable i uniformly at random for both chains to sample. Similarly, choose the HOGWILD! delays $\tilde{\tau}_{i,t}$ to also be the same between the two chains. At time t, let $\tilde{U}_t$ denote the state that would be read by chain X's sampler based on the delays, and similarly let $\tilde{V}_t$ denote the state that would be read by chain Y's sampler. That is,

$$\tilde{U}_{i,t} = X_{i,t-\tilde{\tau}_{i,t}},$$

and similarly,

$$\tilde{V}_{i,t} = Y_{i,t-\tilde{\tau}_{i,t}}.$$

As in the sequential case, we sample the selected variable in both chains using the optimal coupling (of the conditional distributions of the variable to be sampled in both chains) guaranteed by Proposition 1. Iterated over time, this defines a full coupling of the two chains.

We follow the same argument as in the sequential case. First, consider the event that $X_{i,t+1} \neq Y_{i,t+1}$. This event will occur if one of two things happens: either we didn't sample variable i at time t and $X_{i,t} \neq Y_{i,t}$; or we did sample variable i at time t, and the sampled variables were not equal. Since the probability of sampling variable i is $\frac{1}{n}$, and we know the probability that the sampled variables were not equal from Proposition 1, it follows that, by the law of total probability,

$$P\left(X_{i,t+1} \neq Y_{i,t+1}\right) = \left(1-\frac{1}{n}\right)P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{1}{n}\mathbf{E}\left[\left\|\pi_i(\cdot|\tilde{U}_t) - \pi_i(\cdot|\tilde{V}_t)\right\|_{\mathrm{TV}}\right],$$

where $\pi_i(\cdot|X_t)$ denotes the conditional distribution of variable i in π given the values of the other variables in X_t.

Next, we apply the Lemma 3, which gives us

$$\begin{aligned} P\left(X_{i,t+1} \neq Y_{i,t+1}\right) &\leq \left(1-\frac{1}{n}\right)P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n}\max_j P\left(U_{j,t} \neq V_{j,t}\right) \\ &= \left(1-\frac{1}{n}\right)P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right)P\left(X_{j,t-k} \neq Y_{j,t-k}\right). \end{aligned}$$

Now, if we let

$$\phi_t = \max_i P\left(X_{i,t} \neq Y_{i,t}\right),$$

then maximizing the previous expression over i implies that

$$\phi_{t+1} \leq \left(1-\frac{1}{n}\right)\phi_t + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right)\phi_{t-k}.$$

Now, for some constant $r \leq n^{-1}$, let y_t be defined to be the sequence

$$y_t = \exp(-rt).$$

Then, notice that

$$\begin{aligned}\left(1-\frac{1}{n}\right)y_t + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right) y_{t-k} &= \left(1-\frac{1}{n}\right)\exp(-rt) + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right)\exp(-rt+rk) \\ &= \exp(-rt)\left(\left(1-\frac{1}{n}\right) + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right)\exp(rk)\right) \\ &= \exp(-rt)\left(\left(1-\frac{1}{n}\right) + \frac{\alpha}{n}\max_j \mathbf{E}\left[\exp(r\tilde{\tau}_{j,t})\right]\right).\end{aligned}$$

Now, by the convexity of the exponential function,

$$\left(1-\frac{1}{n}\right)y_t + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right) y_{t-k} \leq \exp(-rt)\left(\left(1-\frac{1}{n}\right) + \frac{\alpha}{n}\max_j \left(1 + rn\mathbf{E}\left[\exp\left(\frac{\tilde{\tau}_{j,t}}{n}\right)-1\right]\right)\right).$$

Applying the constraint that

$$\mathbf{E}\left[\exp\left(\frac{\tilde{\tau}_{j,t}}{n}\right)\right] \leq 1 + \frac{\tau^*}{n},$$

we can reduce this to

$$\begin{aligned}\left(1-\frac{1}{n}\right)y_t + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right) y_{t-k} &\leq \exp(-rt)\left(\left(1-\frac{1}{n}\right) + \frac{\alpha}{n}\left(1+r\tau^*\right)\right) \\ &= y_{t+1}\exp(r)\left(1-\frac{1}{n}+\frac{\alpha}{n}+\frac{r\alpha\tau^*}{n}\right) \\ &\leq y_{t+1}\exp(r)\exp\left(-\frac{1}{n}+\frac{\alpha}{n}+\frac{r\alpha\tau^*}{n}\right) \\ &= y_{t+1}\exp\left(\frac{n+\alpha\tau^*}{n}r - \frac{1-\alpha}{n}\right).\end{aligned}$$

Now, we choose r such that the argument to this exponential is zero; that is, we choose

$$r = \frac{1-\alpha}{n+\alpha\tau^*}.$$

Notice that this choice satisfies the earlier assumption that $0 < r \leq n^{-1}$. Using this choice, we can conclude that

$$y_{t+1} \geq \left(1-\frac{1}{n}\right)y_t + \frac{\alpha}{n}\max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t}=k\right) y_{t-k}.$$

Therefore, by Lemma 7,

$$\phi_t \leq y_t = \exp\left(-\frac{1-\alpha}{n+\alpha\tau^*}t\right).$$

This proves the lemma. □

Lemma 6. *Consider any model of* HOGWILD!*-Gibbs sampling on a distribution π with total influence α. Then if for any initial states (X_0, Y_0) we can construct a coupling (X_t, Y_t) such that the process X_t is distributed according to the dynamics of* HOGWILD!*-Gibbs, the process Y_t is distributed according to the dynamics of sequential Gibbs, and for any time t,*

$$\max_i P\left(X_{i,t+1} \neq Y_{i,t+1}\right) \leq \left(1-\frac{1-\alpha}{n}\right)\max_i P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha\tau}{n^2}.$$

As a secondary result, if the chain satisfies Dobrushin's condition ($\alpha < 1$), then for any variable i and any time t,

$$P\left(X_{i,t} \neq Y_{i,t}\right) \leq \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\alpha\tau}{(1-\alpha)n}.$$

Proof of Lemma 6. Define the coupling as follows. Start in state (X_0, Y_0), and at each timestep, choose a single variable $\tilde{I}_t$ uniformly at random for both chains to sample. Then, choose the delays $\tilde{\tau}_{i,t}$ for the HOGWILD! chain X_t. At time t, let $\tilde{U}_t$ denote the state that would be read by chain X's sampler based on the delays. That is,

$$\tilde{U}_{i,t} = X_{i,t-\tilde{\tau}_{i,t}}.$$

As done previously, we sample the selected variable $\tilde{I}_t$ in both chains using the optimal coupling guaranteed by Proposition 1. Iterated over time, this defines a full coupling of the two chains.

We follow a similar argument as in the above lemmas used to bound the mixing time. First, consider the event that $X_{i,t+1} \neq Y_{i,t+1}$. This event will occur if one of two things happens: either we didn't sample variable i at time t and $X_{i,t} \neq Y_{i,t}$; or we did sample variable i at time t, and the sampled variables were not equal. Since the probability of sampling variable i is $\frac{1}{n}$, and we know the probability that the sampled variables were not equal from Proposition 1, it follows that, by the law of total probability,

$$P\left(X_{i,t+1} \neq Y_{i,t+1}\right) = \left(1 - \frac{1}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{1}{n}\mathbf{E}\left[\left\|\pi_i(\cdot|\tilde{U}_t) - \pi_i(\cdot|\tilde{Y}_t)\right\|_{\mathrm{TV}}\right],$$

where $\pi_i(\cdot|X_t)$ denotes the conditional distribution of variable i in π given the values of the other variables in X_t.

Next, we apply the Lemma 3, which gives us

$$\begin{aligned} P\left(X_{i,t+1} \neq Y_{i,t+1}\right) &\leq \left(1 - \frac{1}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n} \max_j P\left(U_{j,t} \neq Y_{j,t}\right) \\ &= \left(1 - \frac{1}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n} \max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t} = k\right) P\left(X_{j,t-k} \neq Y_{j,t}\right). \end{aligned}$$

In order to evaluate this, we notice that the event $X_{j,t-k} \neq Y_{j,t}$ can happen only if either $X_{j,t} \neq Y_{j,t}$ or at some time s, where $t - k \leq s < t$, we sampled variable j (that is, $\tilde{I}_s = j$). Therefore, by the union bound,

$$P\left(X_{j,t-k} \neq Y_{j,t}\right) \leq P\left(X_{j,t} \neq Y_{j,t}\right) + \sum_{s=t-k}^{t-1} P\left(\tilde{I}_s = j\right).$$

Since the probability of sampling variable j at any time is always just $\frac{1}{n}$, we can reduce this to

$$P\left(X_{j,t-k} \neq Y_{j,t}\right) \leq P\left(X_{j,t} \neq Y_{j,t}\right) + \frac{k}{n}.$$

Substituting this into our previous expression produces

$$\begin{aligned} P\left(X_{i,t+1} \neq Y_{i,t+1}\right) &\leq \left(1 - \frac{1}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n} \max_j \sum_{k=0}^{\infty} P\left(\tilde{\tau}_{j,t} = k\right) \left(P\left(X_{j,t} \neq Y_{j,t}\right) + \frac{k}{n}\right) \\ &= \left(1 - \frac{1-\alpha}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha}{n^2} \max_j \mathbf{E}\left[\tilde{\tau}_{j,t}\right] \\ &\leq \left(1 - \frac{1-\alpha}{n}\right) P\left(X_{i,t} \neq Y_{i,t}\right) + \frac{\alpha\tau}{n^2}. \end{aligned}$$

Now, if we let

$$\phi_t = \max_i P\left(X_{i,t} \neq Y_{i,t}\right),$$

then maximizing the previous expression over i implies that

$$\phi_{t+1} \leq \left(1 - \frac{1-\alpha}{n}\right) \phi_t + \frac{\alpha\tau}{n^2}.$$

Subtracting from both sides to identify the fixed point gives us

$$\begin{aligned}\phi_{t+1} - \frac{\alpha\tau}{(1-\alpha)n} &\leq \left(1 - \frac{1-\alpha}{n}\right)\phi_t + \frac{\alpha\tau}{n^2} - \frac{\alpha\tau}{(1-\alpha)n} \\ &= \left(1 - \frac{1-\alpha}{n}\right)\left(\phi_t - \frac{\alpha\tau}{(1-\alpha)n}\right).\end{aligned}$$

Applying this inequality recursively lets us conclude that

$$\begin{aligned}\phi_t - \frac{\alpha\tau}{(1-\alpha)n} &\leq \left(1 - \frac{1-\alpha}{n}\right)^t \left(\phi_0 - \frac{\alpha\tau}{(1-\alpha)n}\right) \\ &\leq \exp\left(-\frac{1-\alpha}{n}t\right),\end{aligned}$$

and so,

$$\phi_t \leq \exp\left(-\frac{1-\alpha}{n}t\right) + \frac{\alpha\tau}{(1-\alpha)n}.$$

This is the desired expression. □

Lemma 7 (Monotonic Sequence Domination Lemma)**.** *Let $x_0, x_1, \ldots$ be a sequence such that, for all t,*

$$x_{t+1} \leq f_t(x_t, x_{t-1}, \ldots, x_0),$$

where f_t is a function that is monotonically increasing in all of its arguments. Then, for any sequence $y_0, y_1, \ldots$, if $x_0 = y_0$ and for all t,

$$y_{t+1} \geq f_t(y_t, y_{t-1}, \ldots, y_0),$$

then for all t,

$$x_t \leq y_t.$$

Proof of Lemma 7. We will approach this by induction. The base case holds by assumption, since $x_0 = y_0$. For the inductive case, if $x_t \leq y_t$ for all $t \leq T$, then

$$x_{T+1} \leq f_T(x_T, x_{T-1}, \ldots, x_0).$$

By monotonicity and the inductive hypothesis,

$$x_{T+1} \leq f_T(y_T, y_{T-1}, \ldots, y_0),$$

and therefore,

$$x_{T+1} \leq y_{T+1}.$$

Applying induction to this proves the lemma. □

Lemma 8. *Consider the model on N variables X_i, for N odd, where each X_i takes on values in $\{-1, 1\}$ and has probability*

$$\pi(X) = \frac{1}{Z_X}\begin{cases} 1 & \text{if } \left|\mathbf{1}^T X\right| = 1 \\ 0 & \text{if } \left|\mathbf{1}^T X\right| > 1 \end{cases}$$

Then Gibbs sampling on this model (assuming that we allow the chain to start only at a state X where $\pi(X) > 0$) has mixing time

$$t_{\text{mix}} = O(n \log n).$$

Proof of Lemma 8. (This lemma contains much of the technical work needed to prove Statement 1. A higher-level motivation for why we are proving this lemma is furnished in the proof of that result.)

Assume that, as we run the chain described in this lemma, we also assign a "color" to each of the variables. All variables with an initial value of 1 start out as black, and all other variables start out as white. Let B_t denote the set of variables that are colored black at any time t, and let S_t denote the sum of all variables that are colored black at that time. We re-color variables according to the following procedure:

1. Whenever we change a variable's value from -1 to 1, if it is colored white, color it black.
2. Whenever we change a variable's value from -1 to 1, if it is already colored black, choose a random variable that had value -1 at time t, and if it is white, color it black.

Note that as a consequence of this result, a variable that is colored white always has value -1.

We will prove the following sub-result by induction on t: given a time t, set B_t, and sum S_t, the values of the variables in B_t are uniformly distributed over the set of possible assignments that are consistent with S_t.

(Base Case.) The base case is straightforward. Since B_0 is just the set of variables that have value 1, there is only one possible assignment that is consistent with S_0: the assignment in which all variables take on the value 1. Since this assignment actually occurs with probability 1, the statement holds.

(Inductive Case.) Assume that the sub-result is true at time t. The sampler chooses a new variable i to sample. One of the following things will happen:

- We don't re-color any variables, or change the values of any variables in B_t. In this case, $B_{t+1} = B_t$ and $S_{t+1} = S_t$. Since there is no change to B or S, all consistent assignments of the black variables are still equiprobable.
- We don't re-color any variables, but we do change the value of some variable in B_t (by changing its value from 1 to -1). Since we sampled the variable i at random, all consistent assignments of the black variables will remain equiprobable.
- We re-color some variable j black. There are two events that can cause this:
  - We could have sampled variable j (that is $i = j$), and changed its value from -1 to 1. This will happen with probability
  $$\frac{1}{N} \cdot \frac{1}{2} = \frac{1}{2N}$$
  - We could have sampled a variable $i \neq j$ that is already colored black, changed its value from -1 to 1, and then chosen variable j at random to color black. Since, at time t, the number of variables with value -1 must be
  $$\frac{N+1}{2},$$
  (since we are about to change a value from -1 to 1), this will happen with probability
  $$\frac{u}{N} \cdot \frac{1}{2} \cdot \frac{2}{N+1} = \frac{u}{N(N+1)}$$
  where u is the number of black-colored variables that have value -1 at time t.

  From this analysis, it follows that, given that we re-colored some variable j black, it will have value -1 with probability
  $$P\left(\text{variable } j \text{ has value } -1\right) = \frac{\frac{u}{N(N+1)}}{\frac{1}{2N} + \frac{u}{N(N+1)}} = \frac{u}{u + \frac{N+1}{2}}.$$
  In particular, at time t, the number of variables that are in B_t is
  $$\frac{N-1}{2} + u,$$

since all variables with value 1 are in B_t, and B_t is stipulated to contain u additional variables with value -1. It follows that at time $t+1$, the number of variables that are in B_t is

$$\frac{N+1}{2} + u,$$

and there will still be u variables in B_{t+1} with value -1. Therefore, the fraction of variables in B_{t+1} that have value -1 will be

$$\frac{u}{u + \frac{N+1}{2}}.$$

Note that this is exactly equal to the probability that variable j will have value -1. Combining this with the inductive hypothesis shows that the consistent states will all remain equiprobable in this case.

Since the consistent states remain equiprobable in all of the possible cases, it follows from the law of total probability that the consistent states are equiprobable in all cases. This shows that the sub-result holds in the inductive case.

We have now showed that given a time t, set B_t, and sum S_t, the values of the variables in B_t are uniformly distributed over the set of possible assignments that are consistent with S_t. This implies that if T_1 is the first time at which the set B_t contains all variables, the value of X_T is are uniformly distributed over all possible states with $\mathbf{1}^T X = 1$.

Now, we performed this construction for a particular polarity of swaps (i.e. focusing on switches from -1 to 1), but by symmetry we could just as easily have used the same construction with the signs of all the variables reversed. If we let T_{-1} be the first time at which the set B_t contains all variables using this reverse-polarity construction, then the value of X_T is uniformly distributed over all possible states with $\mathbf{1}^T X = -1$.

Let T^* be a random variable that is T_1 with probability $\frac{1}{2}$ and T_{-1} with probability $\frac{1}{2}$. It follows that at time T^*, the distribution of X_{T^*} will be π. Therefore, T^* is a strong stationary time for this chain. By the properties of strong stationary times,

$$t_{\text{mix}} \leq 4\mathbf{E}\left[T^*\right].$$

To bound the mixing time, we start by noticing that

$$\mathbf{E}\left[T^*\right] = \frac{1}{2}\mathbf{E}\left[T_1\right] + \frac{1}{2}\mathbf{E}\left[T_{-1}\right] = \mathbf{E}\left[T_1\right].$$

If we let $\bar{T}$ be the first time at which each variable has been set to 1 at least once, then

$$T_1 \leq \bar{T}.$$

Now, if we sample a variable, the probability that we will set it to 1 is (roughly) $\frac{1}{4}$. It follows from the coupon collector's problem bound that the expected amount of time required to set all variables to 1 at least once is

$$\mathbf{E}\left[\bar{T}\right] = O(n \log n).$$

Combining this with the previous inequalities lets us conclude that

$$t_{\text{mix}} = O(n \log n),$$

which proves the lemma. □