Improving Gibbs Sampler Scan Quality with DoGS

Ioannis Mitliagkas [1] **Lester Mackey** [2]

Abstract

The pairwise influence matrix of Dobrushin has long been used as an analytical tool to bound the rate of convergence of Gibbs sampling. In this work, we use Dobrushin influence as the basis of a practical tool to certify and efficiently improve the quality of a Gibbs sampler. Our Dobrushin-optimized Gibbs samplers (DoGS) offer customized variable selection orders for a given sampling budget and variable subset of interest, explicit bounds on total variation distance to stationarity, and certifiable improvements over the standard systematic and uniform random scan Gibbs samplers. In our experiments with joint image segmentation and object recognition, Markov chain Monte Carlo maximum likelihood estimation, and Ising model inference, DoGS consistently deliver higher-quality inferences with significantly smaller sampling budgets than standard Gibbs samplers.

1. Introduction

The Gibbs sampler of Geman and Geman (Geman & Geman, 1984), also known as the *Glauber dynamics* or the *heat-bath algorithm*, is a leading Markov chain Monte Carlo (MCMC) method for approximating expectations unavailable in closed form. First detailed as a technique for restoring degraded images (Geman & Geman, 1984), Gibbs sampling has since found diverse applications in statistical physics (Janke, 2008), stochastic optimization and parameter estimation (Geyer, 1991), and Bayesian inference (Lunn et al., 2000).

The hallmark of any Gibbs sampler is conditional simulation: individual variables are successively simulated from the univariate conditionals of a multivariate target distribution. The principal degree of freedom then is the *scan*, the order in which each variable is sampled. While it is common to employ a *systematic scan*, sweeping through each variable in turn, or a *uniform random scan*, sampling each variable with equal frequency, it is known that non-uniform scans can lead to more accurate inferences both in theory and in practice (Liu et al., 1995; Levine & Casella, 2006). This effect is particularly pronounced when certain variables are of greater inferential interest. Past approaches to optimizing Gibbs sampler scans have been based on asymptotic quality measures approximated with the output of a Markov chain (Levine et al., 2005; Levine & Casella, 2006).

In this work, we propose a computable non-asymptotic scan quality measure based on Dobrushin's notion of variable influence (Dobrushin & Shlosman, 1985). For a given subset of variables, this *Dobrushin variation* bounds the marginal total variation between a target distribution and T steps of Gibbs sampling with a specified scan. More generally, Dobrushin variation bounds a weighted total variation based on user-inputted importance weights for each variable. We couple this quality measure with an efficient procedure for optimizing scan quality by minimizing Dobrushin variation. Our *Dobrushin-optimized Gibbs samplers* (*DoGS*) come equipped with a guaranteed bound on scan quality, are never worse than the standard uniform random and systematic scans, and can be tailored to a target number of sampling steps and a subset of target variables. Moreover, Dobrushin variation can be used to evaluate and compare the quality of any user-specified set of scans prior to running any expensive simulations.

Summary of contributions Building on a coupling formulation traditionally used for the analysis of Gibbs sampler mixing time, we make the following contributions:

- We provide a computable measure of quality for any Gibbs sampler scan and show how it can be used to evaluate and select amongst candidate scans.
- We develop an efficient optimization procedure to improve the quality of an input scan.
- We show how to tailor scans to specific feature classes of a target distribution like expectations depending only on a subset of variables. Our optimization in those cases leads to a dramatic reduction in the number of sampling steps.

[1]Department of Computer Science, Stanford University, Stanford, CA 94305 USA [2]Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142 USA . Correspondence to: Ioannis Mitliagkas <imit@stanford.edu>, Lester Mackey <lmackey@microsoft.com>.

- We demonstrate the advantages of our customized scans for joint image segmentation and object recognition, Markov chain Monte Carlo maximum likelihood estimation, and inference in the Ising model.

The remainder of the paper is organized as follows. Section 2 reviews Gibbs sampling and standard but computationally intractable measures of Gibbs sampler quality. In Section 3, we introduce our scan quality measure and its relationship to total variation. We describe our procedures for selecting high-quality Gibbs sampler scans in Section 4. In Section 5, we apply our techniques to three popular applications of the Gibbs sampler: joint image segmentation and object recognition, MCMC maximum likelihood estimation with intractable gradients, and inference in the Ising model. In each case, we observe substantial improvements in full or marginal total variation over standard Gibbs sampler scans. Section 6 presents our conclusions and discussion of future work.

Notation For any vector v and index i, we let v_{-i} represent the subvector of v with entry v_i removed. We use $\mathrm{diag}(v)$ to represent a square diagonal matrix with the elements of v on the diagonal. The i-th standard basis vector is denoted by e_i, I represents an identity matrix, $\mathbf{1}$ signifies a vector of ones, and $\|C\|$ is the spectral norm of a matrix C. We use $[p]$ to denote the set of all integers from 1 to p.

2. Gibbs sampling and total variation

Consider a target distribution π on a finite p-dimensional state space, $\mathcal{X}^p$. Our inferential goal is to approximate expectations – means, moments, marginals, and more complex function averages, $\mathbb{E}_\pi[f(X)] = \sum_{x\in\mathcal{X}^p} \pi(x)f(x)$ – under π, but we assume that both exact computation and direct sampling from π are prohibitive due to the large number of states, $|\mathcal{X}|^p$. Markov chain Monte Carlo (MCMC) algorithms attempt to skirt this intractability by simulating a sequence of random vectors $X^0, X^1, \ldots, X^T \in \mathcal{X}^p$ from tractable distributions such that expectations over X^T are close to expectations under π. Algorithm 1 summarizes the specific recipe employed by the Gibbs sampler (Geman & Geman, 1984), a leading MCMC algorithm which successively simulates single variables from their tractable conditional distributions.

The principal degree of freedom in a Gibbs sampler is the *scan*, the sequence of p-dimensional probability vectors $q_1, \ldots, q_T$ determining the probability of resampling each variable on each round of Gibbs sampling. Typically one selects between the uniform random scan, $q_t = (1/p, \ldots, 1/p)$ for all t, where variable indices are selected uniformly at random on each round and the systematic scan, $q_t = e_{(t \bmod p)+1}$ for each t, which repeatedly cycles through each variable in turn. However, non-uniform scans are known to lead to better approximations (Liu et al., 1995; Levine & Casella, 2006), motivating the need for practical procedures for evaluating and improving Gibbs sampler scans.

Algorithm 1 Gibbs sampling (Geman & Geman, 1984)

input Scan $(q_t)_{t=1}^T$; starting distribution μ; single-variable conditionals of target distribution, $\pi(\cdot|X_{-i})$
  $X^0 \sim \mu$ (Sample from starting distribution)
  for t in $1, 2, \ldots, T$ **do**
    Sample variable index to update using scan: $i_t \sim q_t$
    Sample $X^t_{i_t} \sim \pi(\cdot|X^{t-1}_{-i_t})$ from its conditional
    Copy remaining variables: $X^t_{-i_t} = X^{t-1}_{-i_t}$
  end for
output Sample sequence $(X^t)_{t=0}^T$

Let $P \in \mathbb{R}^{|\mathcal{X}|^p \times |\mathcal{X}|^p}$ denote the transition probability matrix of the Gibbs sampler Markov chain $(X^t)_{t=0}^T$, and, for each step t, let $\pi_t \triangleq {P^t}^\top \mu$ be the distribution of X^t. The quality of a T-step Gibbs sampler and its scan is typically measured in terms of total variation distance between π_t and the target distribution π:

Definition 1. *The total variation distance between probability measures μ and ν is the maximum difference in expectations over all $[0,1]$-valued functions,*

$$\|\mu - \nu\|_{TV} \triangleq \sup_{f:\mathcal{X}^p \to [0,1]} |\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(Y)]|.$$

The following lemma gives a coupling characterization of total variation.

Lemma 2 (Prop. 4.7 of (Levin et al., 2009))**.**

$$\|\mu - \nu\|_{TV} = \inf_{(X,Y): X\sim\mu, Y\sim\nu} \mathbb{P}(X \neq Y)$$

As we will see in Section 3, this characterization of total variation plays an important role in a classic style of analysis for Gibbs sampling.

2.1. Marginal total variation

While we typically sample all p variables in the process of Gibbs sampling, it is common for some variables to be of greater interest than others. For example, when modeling a large particle system, we may be interested principally in the behavior in local region of the system; likewise, when automatically segmenting an image into its component parts, a particular region of the image, like the area surrounding a face, is often of primary interest. In these cases, it is more natural to consider a marginal total variation that measures the discrepancy in expectation over only those variables of interest.

Definition 3 (Marginal total variation)**.** *The marginal total variation between probability measures μ and ν on a subset of variables $S \in [p]$ is the maximum difference in expectations over all $[0,1]$-valued functions of $X\big|_S$, the restriction of X to the coordinates in S:*

$$\|\mu - \nu\|_{S,TV} \triangleq \sup_{f:\mathcal{X}^{|S|}\to[0,1]} \left|\mathbb{E}_\mu\left[f\left(X\big|_S\right)\right] - \mathbb{E}_\nu\left[f\left(X\big|_S\right)\right]\right|.$$

More generally, we will seek to control an arbitrary user-defined weighted total variation that assigns an independent non-negative weight to each coordinate and hence controls approximation error for functions with varying sensitivities in each variable.

Definition 4 (d-Lipschitz feature)**.** *Let $f : \mathcal{X}^p \to \mathbb{R}$ be a function on p variables with the property*

$$\sup_{X\in\mathcal{X}^p, z\in\mathcal{X}} |f(X_1,\dots,X_p) - f(X_1,\dots,X_{i-1},z,X_{i+1},\dots,X_p)| \le d_i$$

These are Lipschitz constants for each coordinate under the Hamming metric. We call this kind of function a $(d_i,\dots,d_p)$-Lipschitz feature.

For example, every function with range $[0,1]$ is a $\mathbf{1}$-Lipschitz feature, and the value of the first variable, $x \mapsto x_1$, is an e_1-Lipschitz feature. This definition leads to a measure of sample quality specific to a d-Lipschitz class of features.

Definition 5 (d-weighted total variation)**.** *The d-weighted total variation between probability measures μ and ν is the maximum difference in expectations across d-Lipschitz functions:*

$$\|\mu - \nu\|_{d,\mathrm{TV}} \triangleq \sup_{d-\mathrm{Lipschitz}\, f} |\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(Y)]|$$

3. Measuring scan quality with Dobrushin variation

Direct computation of total variation measures is expensive. Common methods generate diagnostic sequences that often increase running time by a whole order of magnitude (cf. Discussion in (Brooks et al., 1997)). We provide an efficiently computable upper bound on the weighted total variation of Definition 5. The quality measure we introduce, as well as the algorithms we suggest for finding good selection distributions and sequences are based on a coupling analysis due to Dobrushin (Dobrushin & Shlosman, 1985).

These results critically depend on the *influence matrix* of the target, a quantity central to past convergence analyses of Gibbs sampling, and on Dobrushin's condition, both defined in this section. Past work has used the influence matrix as a tool for analysis, to show that systematic and uniform random scan Gibbs samplers *mix* quickly – that is, provide estimates that converge rapidly to the target expectations – when the $\|C\| < 1$ for any matrix norm (Dyer et al., 2009). Here, we use the influence matrix as a computational tool to compute and improve the quality of Gibbs sampling.

Consider any sequence of variables $(Y^t)_{t=0}^T$ with $Y^t \sim \pi$ for all t, and define the marginal coupling probability $p_t(i) \triangleq \mathbb{P}(X_i^t \neq Y_i^t)$. The following lemma, proved in Appendix A.1, shows that weighted total variation is controlled by such marginal coupling probabilities.

Lemma 6 (Marginal coupling controls weighted TV)**.** *Suppose that $X \sim \mu$ and $Y \sim \nu$ for probability measures μ and ν on $\mathcal{X}^p$. Then, for any nonnegative weight vector $d \in \mathbb{R}^p$,*

$$\|\mu - \nu\|_{d,\mathrm{TV}} \le \sum_i d(i)\mathbb{P}(X_i \neq Y_i).$$

Dobrushin's approach to controlling the marginal coupling probabilities p_t uses influence, a measure of how much changing variable j can affect the conditional of variable i.

Definition 7 (Dobrushin influence matrix)**.** *The influence of variable j on i is given by*

$$C_{ij} \triangleq \max_{(X,Y)\in N_j} \|\pi(\cdot|X_{-i}) - \pi(\cdot|Y_{-i})\|_{TV} \qquad (1)$$

where $(X,Y) \in N_j$ signifies $X_l = Y_l$ for all $l \neq j$.

Now we have all the ingredients to formulate our quality measure for any scan and give guarantees on its bias-controlling powers.

Definition 8 (Dobrushin variation)**.** *The* Dobrushin variation *of scan $(q_t)_{t=1}^T$ for a d-Lipschitz feature and an entrywise upper bound $\bar{C}$ on the Dobrushin influence (1) is*

$$\mathcal{V}(q_1,\dots,q_T; d, \bar{C}) \triangleq d^\top B(q_T)\cdots B(q_1)\mathbf{1} \qquad (2)$$

for $B(q) \triangleq (I - \mathrm{diag}(q)(I - \bar{C}))$.

The next result, proved in Appendix A.2, yields a model- and scan-specific guarantee on the weighted total variation quality of a Gibbs sampler in terms of the efficiently computable Dobrushin variation $\mathcal{V}$.

Theorem 9 (Dobrushin variation controls weighted TV)**.** *Suppose that π_T is the distribution of the T-th step of a Gibbs sampler with scan $(q_t)_{t=1}^T$. Then, for any nonnegative weight vector d and entrywise upper bound $\bar{C}$ on the Dobrushin influence (1),*

$$\|\pi_T - \pi\|_{d,\mathrm{TV}} \le \mathcal{V}\big((q_t)_{t=1}^T; d, \bar{C}\big).$$

4. Improving scan quality with DoGS

In this section, we present an efficient algorithm for improving the quality of any Gibbs sampler scan by optimizing

Algorithm 2 DoGS: Scan selection via coordinate descent

input Scan $(q_\tau)_{\tau=1}^T$; variable weights d; influence entrywise upper bound $\bar{C}$; (optional) target accuracy ϵ.
  Initialize $d_T \leftarrow d$
  for t in $T, T-1, \dots, 1$ **do**
    If $d_t^\top B(q_t)\cdots B(q_1)\mathbf{1} \le \epsilon$, then **break**
    $r_t \leftarrow \text{argmax}_i\, d_t(i) e_i^\top (I-\bar{C})B(q_{t-1})\cdots B(q_1)\mathbf{1}$
    $d_{t-1}^\top \leftarrow d_t^\top (I - e_{r_t} e_{r_t}^\top (I-\bar{C}))$
  end for
output Optimized scan $(q_\tau)_{\tau=1}^{t-1}, (r_\tau)_{\tau=t}^T$

Dobrushin variation (Definition 8). We will refer to the resulting customized Gibbs samplers as *Dobrushin-optimized Gibbs samplers* or *DoGS* for short. Algorithm 2 optimizes Dobrushin variation using coordinate descent, with the selection distribution q_t for each time step serving as a coordinate.

A user can initialize DoGS with any baseline scan, including the systematic or uniform random scan, and the resulting customized scan is guaranteed to have the same or better Dobrushin variation. Moreover, DoGS scans will always be d-ergodic (i.e., $\|\pi_T - \pi\|_{d,\text{TV}} \to 0$ as $T \to \infty$) when initialized with a systematic or uniform random scan and $\|\bar{C}\| < 1$. This follows from the following proposition, which shows that Dobrushin variation —and hence the d-weighted total variation by Theorem 9—goes to 0 under these conditions and standard scans.

Proposition 10. *Suppose that $\bar{C}$ is an entrywise upper bound on the Dobrushin influence and $(q_t)_{t=1}^T$ is a systematic or uniform random scan. If $\|\bar{C}\| < 1$, then, for any nonnegative weight vector d, the Dobrushin variation vanishes as the chain length T increases,*

$$\lim_{T\to\infty} \mathcal{V}(q_1, \dots, q_T; d, \bar{C}) = 0.$$

The proof relies on arguments in (Hayes, 2006) and is outlined in Appendix A.3.

Since Dobrushin variation is linear in each q_t, each coordinate optimization in Algorithm 2 (in the absence of ties) selects a degenerate distribution, a single coordinate, yielding a fully deterministic scan. The complete algorithm can be implemented to run in time $O(pT)$ for deterministic input scans and $O(p^2T)$ for random input scans by precomputing the vectors $B(q_t)\cdots B(q_1)\mathbf{1}$ or $(I-\bar{C})B(q_t)\cdots B(q_1)\mathbf{1}$ for all t, at a cost of $O(pT)$ memory.

4.1. Bounding influence

An essential input to our algorithms is the entrywise upper bound $\bar{C}$ on the influence matrix (1). Fortunately, Liu & Domke (2014) showed that influence can be bounded in a straightforward manner for any pairwise Markov random field (MRFs) target,

$$\pi(X) \propto \exp\Big(\textstyle\sum_{i,j}\sum_{a,b\in\mathcal{X}} \theta_{ab}^{ij}\, \mathbb{I}[X_i = a, X_j = b]\Big). \quad (3)$$

Theorem 11 (Pairwise MRF influence (Liu & Domke, 2014, Lems. 10, 11))**.** *Using the shorthand $\sigma(s) \triangleq \frac{1}{1+e^{-s}}$, the influence (1) of the target π in (3) satisfies*

$$C_{ij} \le \max_{x_j, y_j} |2\sigma(\tfrac{1}{2}\max_{a,b}(\theta_{ax_j}^{ij} - \theta_{ay_j}^{ij}) - (\theta_{bx_j}^{ij} - \theta_{by_j}^{ij})) - 1|.$$

Pairwise MRFs with binary variables $X_i \in \{-1, 1\}$ are especially common in statistical physics and computer vision. A general parameterization for binary pairwise MRFs is given by

$$\pi(X) \propto \exp\Big(\textstyle\sum_{i\neq j}\theta_{ij}X_iX_j + \sum_i \theta_i X_i\Big), \quad (4)$$

and our next theorem, proved in Appendix A.4, leverages the strength of the singleton parameters θ_i to provide a tighter bound on the influence of these targets.

Theorem 12 (Binary pairwise influence)**.** *The influence (1) of the target π in (4) satisfies*

$$C_{ij} \le \frac{|\exp(2\theta_{ij}) - \exp(-2\theta_{ij})|b^*}{(1 + b^*\exp(2\theta_{ij}))(1 + b^*\exp(-2\theta_{ij}))}$$

for $b^* = \max(e^{-2\sum_{k\neq j}\theta_{ik} - 2\theta_i}, \min[e^{2\sum_{k\neq j}\theta_{ik} - 2\theta_i}, 1])$.

Theorem 12 in fact provides an exact computation of the Dobrushin influence C_{ij} whenever $b^* \neq 1$. The only approximation comes from the fact that the value $b^* = 1$ may not belong to the set $\mathcal{B} = \{e^{2\sum_{k\neq j}\theta_{ik}X_k - 2\theta_i} \mid X \in \{-1,1\}^p\}$. An exact computation of C_{ij} would replace the cutoff of 1 with its closest approximation in $\mathcal{B}$.

Next we consider a family of *generalized Potts models*, pairwise Markov random fields with non-negative interaction parameters θ_{ij} and general discrete variables $X_i \in \mathcal{X}$:

$$\pi(X) \propto \exp(\textstyle\sum_i\sum_j \theta_{ij}\mathbb{I}\{X_i = X_j\} + \sum_i\sum_{a\in\mathcal{X}}\theta_{i,a}\mathbb{I}\{X_i = a\}). \quad (5)$$

Theorem 13 bounds the influence for this class of distributions commonly employed in computer vision.

Theorem 13 (Generalized Potts influence)**.** *For the model in (5) with pairwise $\theta_{ij} \ge 0$, the influence (1) satisfies*

$$C_{ij} \le \frac{\exp\Big(\max_{a\in\mathcal{X}}\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big)(\exp(\theta_{ij}) - 1)}{\Big[\begin{array}{l}\exp(\max_{a\in\mathcal{X}}\theta_{i,a})(\exp\Big(\sum_{k\neq j}\theta_{ik}\Big) - 1) \\ \quad + \exp(\min_{b\in\mathcal{X}}\theta_{i,b})(\exp(\theta_{ij}) - 1) + \sum_{c\in\mathcal{X}}\exp(\theta_{i,c})\end{array}\Big]}.$$

The proof can be found in Appendix A.5.

4.2. Related Work

In related work, Latuszynski et al. (2013) recently analyzed an abstract class of adaptive Gibbs samplers parameterized by an arbitrary scan selection rule. However, as noted in their Rem. 5.13, no explicit scan selection rules were provided in that paper. The only prior concrete scan selection rules of which we are aware are the Minimax Adaptive Scans with asymptotic variance or convergence rate objective functions (Levine & Casella, 2006). Unless some substantial approximation is made, it is unclear how to implement these procedures when the target distribution of interest is not Gaussian.

Levine & Casella (2006) approximate these Minimax Adaptive Scans for specific mixture models by considering single ad hoc features of interest; the approach has many hyperparameters to tune including the order of the Taylor expansion approximation, which sample points are used to approximate asymptotic quantities online, and the frequency of adaptive updating. Our proposed quality measure, Dobrushin variation, requires no approximation or tuning and can be viewed as a practical non-asymptotic objective function for the abstract scan selection framework of Levine and Casella. In the spirit of (Lacoste-Julien et al., 2011), DoGS can also be viewed as an approximate inference scheme calibrated for downstream inferential tasks depending only on subsets of variables.

Levine et al. (2005) employ the Minimax Adaptive Scans of Levine and Casella by finding the mode of their target distribution using EM and then approximating the distribution by a Gaussian. They report that this approach to scan selection introduces substantial computational overhead (10 minutes of computation for an Ising model with 64 variables). As we will see in Section 5, the overhead of DoGS scan selection is manageable (15 seconds of computation for an Ising model with 1 million variables) and outweighed by the increase in scan quality and sampling speed.

5. Experiments

In this section, we demonstrate how our proposed scan quality measure and efficient optimization schemes can be used to both evaluate and improve Gibbs sampler scans when either the full distribution or a marginal distribution is of principal interest. For all experiments with binary MRFs, we have use Theorem 12 to produce the Dobrushin influence bound $\bar{C}$. On all ensuing plots, the numbers in the legend give the best guarantee achieved for each algorithm plotted. Due to space constraints, we display only one representative plot per experiment; the analogous plots from independent replicates of each experiment can be found in Appendix B.

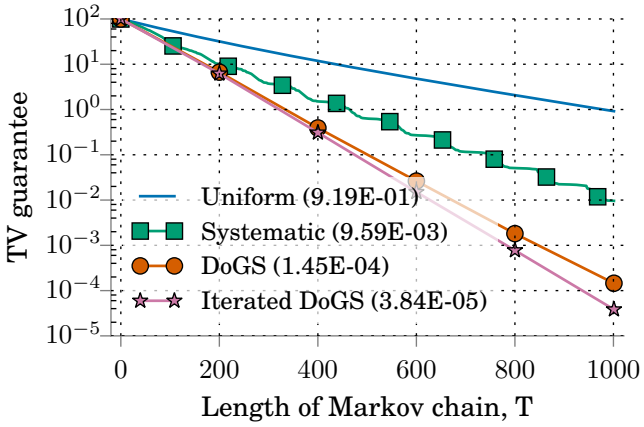

Figure 1. TV guarantees provided Dobrushin variation for various Gibbs sampler scans on a 10×10 non-toroidal Ising model with random parameters (see Section 5.1). DoGS is initialized with the systematic scan.

5.1. Evaluating and optimizing Gibbs sampler scans

In our first experiment, we illustrate how Dobrushin variation can be used to select between standard scans and how DoGS can be used to efficiently improve upon standard scan quality when total variation quality is of interest. We remind the reader that both scan evaluation and scan selection are performed offline prior to any expensive simulation from the Gibbs sampler. Our target is a 10×10 Ising model arranged in a two-dimensional lattice, a standard model of ferromagnetism in statistical physics. In the notation of (4), we draw the unary parameters θ_i uniformly at random from $\{0, 1\}$, and the interaction parameters uniformly at random: $\theta_{ij} \sim \text{Uniform}([0, 0.25])$.

Figure 1 compares, as a function of the number of steps T, the total variation guarantee provided by Dobrushin variation (see Theorem 9) for the standard systematic and uniform random scans. We see that the systematic scan, which traverses variables in row major order, obtains a significantly better TV guarantee than its uniform random counterpart for all sampling budgets T. Hence, the systematic scan would be our standard scan of choice for this target. DoGS (Algorithm 2) initialized with the systematic scan further improves the systematic scan guarantee by two orders of magnitude. Iterating Algorithm 2 on its own scan output until convergence ("Iterated DoGS" in Figure 1) provides an additional improvement . However, since we consistently find that the bulk of the improvement is obtained with a single run of Algorithm 2, non-iterated DoGS remains our recommended default recipe for quickly improving scan quality.

Note that since our TV guarantee is an upper bound provided by the exact computation of Dobrushin variation, the actual

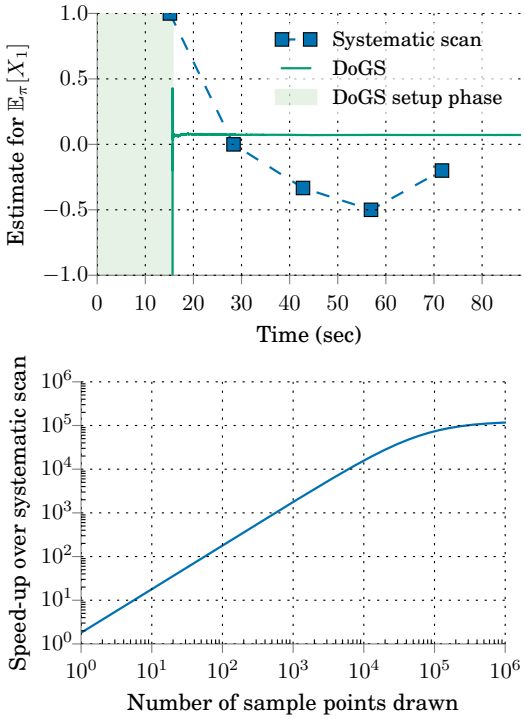

Figure 2. (top) Estimate of target, $\mathbb{E}_\pi[X_1]$, versus wall-clock time for a standard row-major-order systematic scan and a DoGS optimized sequence on an Ising model with 1 million variables (see Section 5.2). By symmetry $\mathbb{E}_\pi[X_1] = 0$. (bottom) The end-to-end speedup of DoGS over systematic scan, including setup and optimization time, as a function of the number of sample points we draw.

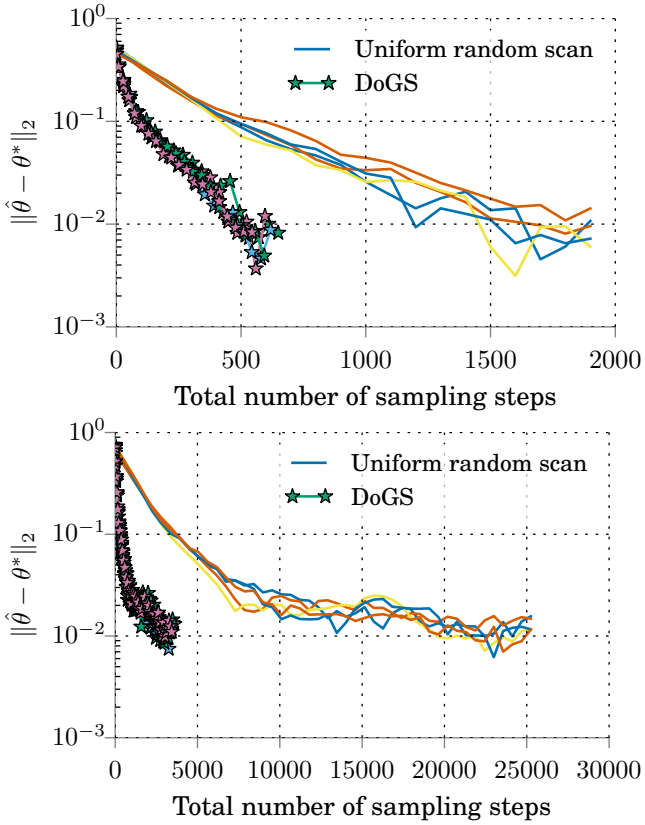

Figure 3. Comparison of parameter estimation error in MCMC maximum likelihood estimation of the 3×3 (top) and a 4×4 (bottom) Ising models of Domke (2015). Each MCMC gradient estimate is obtained either from the uniform random scan suggested by Domke or from DoGS initialized with the uniform random scan, using Algorithm 2 to achieve a target total variation of 0.01 (see Section 5.3). Five runs are shown in each case.

gains in TV may differ from the gains in Dobrushin variation. In practice and as evidenced in Section 5.4, we find that the actual gains in (marginal) TV over standard scans are typically larger than the Dobrushin variation gains.

5.2. End-to-end wall clock time performance

In this experiment, we demonstrate that using DoGS to optimize can result to dramatic speed-ups. This is particularly true for large models and in cases where we draw many samples from the same model. The setting is the same as in the previous experiment, with the exception of model size: here we simulate a $1K \times 1K$ Ising model, with 1 million variables in total. We target a single marginal x_1, and use the samples drawn from a systematic scan that performs two full passes and a DoGS-optimized sequence, to estimate its expectation. To find the DoGS we start from the systematic scan and use a doubling scheme: (i) we start by feeding the first 2 steps of the scan into Algorithm 2; (ii) if DoGS guarantees the same Dobrushin variation as the systematic scan, or better, we keep it—otherwise we double the size of the seed sequence and go back to step (i). Figure 2 reports the estimate of a marginal expectation versus as a function of computation time, including the 15s setup time for DoGS on the 1 million Ising variable model. For comparison, Levine et al. (2005) report 10 minutes of setup time for their adaptive Gibbs scans when processing a 64 variable Ising model. The bottom plot of Figure 2 uses the average measured time for a single step[1], the measured setup time for DoGS and the size of the two scan sequences ($2M$ steps for systematic, 16 steps for DoGS) to give an estimate of the speedup as a function of the number of samples we draw. Additional timing experiments are deferred to Appendix B.2.

5.3. Accelerated MCMC maximum likelihood estimation

Here we demonstrate how DoGS can be used to accelerate MCMC maximum likelihood estimation, while providing

[1]Each Gibbs step took 12.65μs on a 2015 Macbook Pro.

guarantees on parameter estimation quality. We replicate the Ising model maximum likelihood estimation experiment of (Domke, 2015, Sec. 6) and show how we can provide the same level of accuracy faster. We run the uniform random scan parametrized according to (Domke, 2015) and observe the estimation quality it achieves. We then use this value as the target accuracy in Algorithm 2 to generate a DoGS sequence, which we use on the same random models used in the random scan runs. We notice that DoGS attains the same level of parameter estimation accuracy in many fewer sampling steps.

Figure 3 shows the two approaches for two models; each experiment is repeated five times. The original algorithm marked as 'Uniform random scan', achieves a parameter estimation error of about 0.01 . For our method, we use the current parameter values θ_k at every iteration to find the shortest sampling sequence that achieves bias of < 0.01. We start with a systematic scan, which is fed into Algorithm 2, along with $\epsilon = 0.01$. We use the result of Theorem 12 to bound influence and report 5 runs for each case. The results show that DoGS consistently achieves the desired parameter accuracy much more quickly than standard Gibbs.

5.4. Customized scans for fast marginal mixing

In this section we demonstrate how DoGS can be used to dramatically speed up marginal inference while providing model-dependent guarantees. We use a 40×40 non-toroidal Ising model and set our feature to be the top left variable, i.e. $d = [1, 0, 0, \ldots, 0]^\top$. Figure 4 compares guarantees for a uniform random scan and a systematic scan; we also see how we can further improve the total variation guarantees by feeding a systematic scan into Algorithm 2. Again we see that a single run of Algorithm 2 yields the bulk of the improvement and iterated applications only provide small further benefits. For the DoGS sequence, the figure also shows a histogram of the distance of sampled variables from the target variable, X_1, at the top left corner of the grid.

Figure 5 shows that optimizing our objective actually improves performance by reducing the marginal bias much more quickly than systematic scan. For completeness, we include more experiments on a toroidal Ising model in Appendix B.3.

5.5. Targeted image segmentation and object recognition

The Markov field aspect model (MFAM) of Verbeek & Triggs (2007) is a generative model for images designed to automatically divide an image into its constituent parts (image segmentation) and label each part with its semantic object class (object recognition). For each test image k, the MFAM extracts a discrete feature descriptor from each image patch i, assigns a latent object class label $X_i \in \mathcal{X}$ to

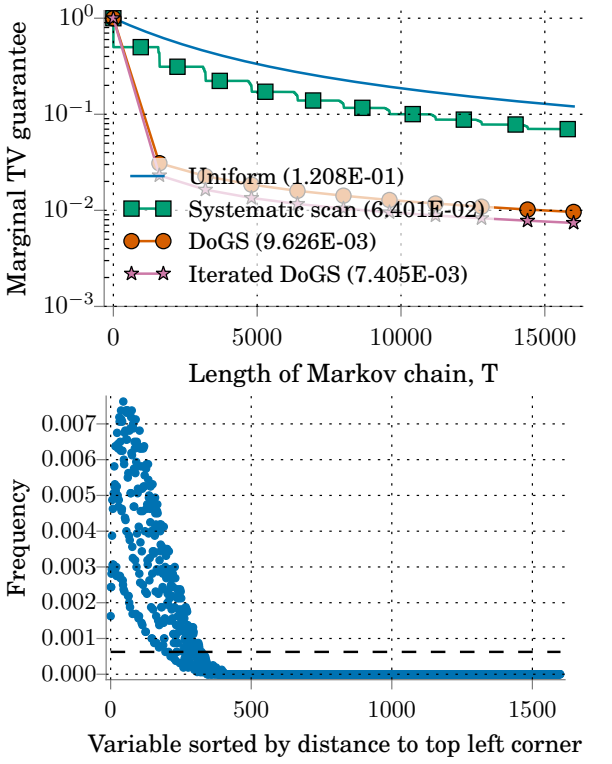

Figure 4. (top) Marginal TV guarantees provided by Dobrushin variation for various Gibbs sampler scans when targeting the top left corner variable on a 40×40 non-toroidal Ising model with $\theta_{ij} \approx 1/3.915$ (see Section 5.4). DoGS is initialized with the systematic scan. (bottom) Frequency with which each variable is sampled in the DoGS sequence of length 16000, sorted by Manhattan distance to target variable.

each patch, and induces the posterior distribution

$$\pi(X|y;k) \propto \exp\Big(\sum_{(i,j)\text{ spatial neighbors}} \sigma \mathbb{I}\{X_i = X_j\} + \sum_i \log\Big(\sum_{a\in\mathcal{X}} \theta_{k,a}\beta_{a,y_i}\mathbb{I}\{X_i = a\}\Big)\Big), \quad (6)$$

over the configuration of patch levels X. When the Potts parameter $\sigma = 0$, this model reduces to probabilistic latent semantic analysis (PLSA) (Hofmann, 2001), while a positive value of σ encourages nearby patches to belong to similar classes. Using the Microsoft Research Cambridge (MSRC) pixel-wise labeled image database v1[2], we follow the weakly supervised setup of Verbeek & Triggs (2007) to fit the PLSA parameters θ and β to a training set of images and then, for each test image k, use Gibbs sampling to generate patch label configurations X targeting the MFAM posterior (6) with $\sigma = 0.48$. We generate a segmentation by assigning each patch the most frequent label encountered during Gibbs sampling and evaluate the accuracy of this labeling using the Hamming error described in (Verbeek & Triggs, 2007). This experiment is repeated over 20 indepen-

[2]http://research.microsoft.com/vision/cambridge/recognition/

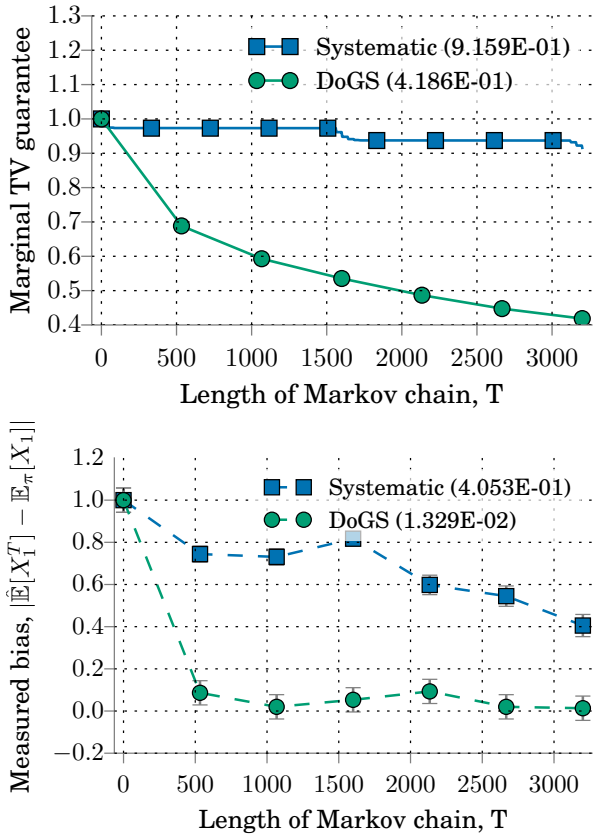

Figure 5. (top) Marginal TV guarantees provided by Dobrushin variation for systematic scan and DoGS initialized with systematic scan when targeting the top left corner variable on a 40×40 toroidal Ising model with $\theta_{ij} = 0.25$ (see Section 5.4). (bottom) Measured bias and standard errors from 300 independent samples of X_1^T.

dently generated 90% training / 10% test partitions of the 240 image dataset.

We select our DoGS scan to target a 12×8 marginal patch rectangle at the center of each image (see Figure 6) and compare its segmentation accuracy and efficiency with that of a standard systematic scan of length $T = 620$. We initialize DoGS with the systematic scan, the influence bound $\bar{C}$ of Theorem 11, and a target accuracy ϵ equal to the marginal Dobrushin variation guarantee of the systematic scan. In 11.5ms, the doubling scheme described in Section 5.2 produced a DoGS sequence of length 110 achieving the Dobrushin variation guarantee ϵ on marginal TV. Figure 7 shows that DoGS achieves a slightly better average Hamming error than systematic scan using a $5.5\times$ shorter sequence. Systematic scan takes 1.2s to resample each variable of interest, while DoGS consumes 0.37s. Moreover, the 11.5ms DoGS scan selection was performed only once and then used to segment all test images. For each chain, X^0 was initialized to the maximum a posteriori patch labeling

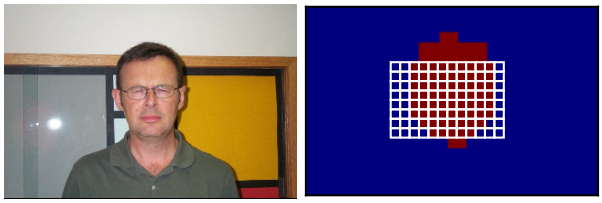

Figure 6. (left) Example test image from MSRC dataset. (right) Segmentation produced by DoGS Markov field aspect model targeting the center region outlined in white (see Section 5.5).

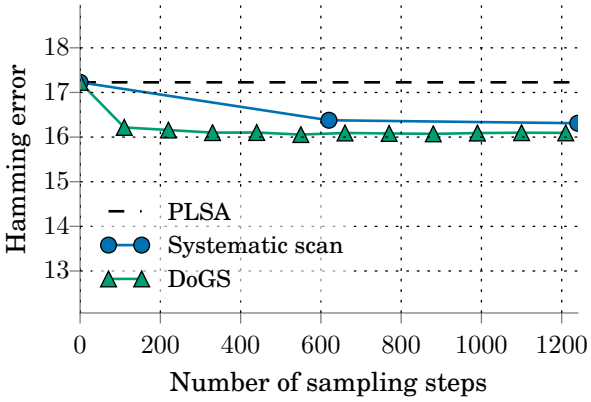

Figure 7. Average test image segmentation error under the Markov field aspect model of Section 5.5. PLSA represents the maximum a posteriori patch labeling under the MFAM (6) with $\sigma = 0$. Errors are averaged over 20 MSRC test sets of 24 images.

under the PLSA model (obtained by setting $\sigma = 0$ in the MFAM).

6. Discussion

We introduced a practical quality measure – Dobrushin variation – for evaluating and comparing existing Gibbs sampler scans and efficient procedures – DoGS – for developing customized fast-mixing scans tailored to marginals or distributional features of interest. We deployed DoGS for three common Gibbs sampler applications – joint image segmentation and object recognition, MCMC maximum likelihood estimation, and Ising model inference – and in each case achieved higher quality inferences with significantly smaller sampling budgets than standard Gibbs samplers. In the future, we aim to enlist DoGS for additional applications in computer vision and natural language processing, extend the reach of DoGS to models containing continuous variables, and integrate DoGS into large inference engines built atop Gibbs sampling.

References

Brooks, Steve P, Dellaportas, Petros, and Roberts, Gareth O. An approach to diagnosing total variation convergence of MCMC algorithms. *Journal of Computational and Graphical Statistics*, 6(3):251–265, 1997.

De Sa, Christopher, Olukotun, Kunle, and Ré, Christopher. Ensuring rapid mixing and low bias for asynchronous Gibbs sampling. *arXiv preprint arXiv:1602.07415*, 2016.

Dobrushin, Roland Lvovich and Shlosman, Senya B. Constructive criterion for the uniqueness of Gibbs field. In *Statistical physics and dynamical systems*, pp. 347–370. Springer, 1985.

Domke, Justin. Maximum likelihood learning with arbitrary treewidth via fast-mixing parameter sets. In *Advances in Neural Information Processing Systems*, pp. 874–882, 2015.

Dyer, Martin, Goldberg, Leslie Ann, and Jerrum, Mark. Matrix norms and rapid mixing for spin systems. *The Annals of Applied Probability*, pp. 71–107, 2009.

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

Geyer, C. J. Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics: Proc. 23rd Symp. Interface*, pp. 156–163, 1991.

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.

Hofmann, Thomas. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

Janke, Wolfhard. Monte Carlo methods in classical statistical physics. In *Computational Many-Particle Physics*, pp. 79–140. Springer, 2008.

Lacoste-Julien, Simon, Huszár, Ferenc, and Ghahramani, Zoubin. Approximate inference for the loss-calibrated bayesian. In *AISTATS*, pp. 416–424, 2011.

Latuszynski, Krzysztof, Roberts, Gareth O., and Rosenthal, Jeffrey S. Adaptive Gibbs samplers and related MCMC methods. *Ann. Appl. Probab.*, 23(1):66–98, 02 2013. doi: 10.1214/11-AAP806. URL http://dx.doi.org/10.1214/11-AAP806.

Levine, R. A. and Casella, G. Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100, 2006.

Levine, Richard A, Yu, Zhaoxia, Hanley, William G, and Nitao, John J. Implementing random scan Gibbs samplers. *Computational Statistics*, 20(1):177–196, 2005.

Liu, Jun S, Wong, Wing H, and Kong, Augustine. Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 157–169, 1995.

Liu, Xianghang and Domke, Justin. Projecting markov random field parameters for fast mixing. In *Advances in Neural Information Processing Systems*, pp. 1377–1385, 2014.

Lunn, David J, Thomas, Andrew, Best, Nicky, and Spiegelhalter, David. WinBUGS-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4): 325–337, 2000.

Verbeek, Jakob and Triggs, Bill. Region classification with markov field aspect models. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8. IEEE, 2007.

A. Proofs

A.1. Proof of Lemma 6

Let $X \sim \mu$ and $Y \sim \nu$. Now define the sequence $(Z_i)_{i=0}^p$, such that $Z_0 \triangleq X$, $Z_p \triangleq Y$ and Z_{i-1}, Z_i can only differ on element i. By Definition 5 and the compactness of the space of d-Lipschitz features, there exists d-Lipschitz f such that

$$\|\mu - \nu\|_{d,\mathrm{TV}} = |\mathbb{E}[f(X) - f(Y)]|.$$

Then using triangle inequality,

$$\begin{aligned}
|\mathbb{E}[f(X) - f(Y)]| \leq& \left|\mathbb{E}[\sum_{i=1}^{p} f(Z_{i-1}) - f(Z_i)]\right| && (7)\\
\leq& \sum_{i=1}^{p} |\mathbb{E}[f(Z_{i-1}) - f(Z_i)]| && (8)\\
\leq& \sum_{i=1}^{p} \mathbb{P}(X_i \neq Y_i) d_i && (9)\\
=& d^\top p_t && (10)
\end{aligned}$$

A.2. Proof of Theorem 9

First we state a useful result due to Dobrushin & Shlosman (1985).

Lemma 14 ((Dobrushin & Shlosman, 1985), similar arguments can be found in Theorem 6 of (Hayes, 2006) and in (De Sa et al., 2016))**.** *Consider the marginal coupling probability $p_t(i) \triangleq \mathbb{P}(X_i^t \neq Y_i^t)$ and influence matrix, C, we defined in Section 3 and an arbitrary scan sequence $(i_t)_{t=1}^{t=T}$. Then, an application of the law of total probability yields the following bound on the marginal coupling probabilities.*

$$p_t(i) \leq \begin{cases} \sum_{j \neq i} C_{ij} p_{t-1}(j), & i_t = i \\ p_{t-1}(i), & \text{o.w.} \end{cases} \tag{11}$$

and $p_0(i) \leq 1$ for all i.

Proof of Theorem 9 At each time step, $i_t = i$ with probability $q_t(i)$. Let $z_i(t) \triangleq \mathbb{I}\{i_t = i\}$ and $Z(t)$ denote the diagonal matrix with $Z_{ii}(t) = z_i(t)$. Note that $\mathbb{E}Z(t) = \mathrm{diag}(q_t)$. Now, from Lemma 6, and using Lemma 14,

$$\begin{aligned}
&|\mathbb{E}f(X^T) - \mathbb{E}f(Y^T)|\\
&\quad\leq d^\top p_T = \sum_i d_i p_i(T)\\
&\quad\leq \sum_i d_i \left(z_i(T) \sum_{j \neq i} C_{ij} p_j(T-1) + (1 - z_i(T)) p_i(T-1) \right)\\
&\quad= d^\top (I - Z_T(I - C)) p_{T-1}.
\end{aligned}$$

Now, taking an expectation over the randomness of sampling (random variables i_t), we get

$$\mathbb{E}|\mathbb{E}f(X^T) - \mathbb{E}f(Y^T)| \leq d^\top (I - \mathrm{diag}(q_T)(I - C)) \mathbb{E}p_{T-1} \leq d^\top B(q_T) \cdots B(q_1)\mathbf{1} = \mathcal{V}(q_1, \ldots, q_T; d, C) \leq \mathcal{V}(q_1, \ldots, q_T; d, \bar{C}),$$

where we used the fact that p_0 is a vector of probabilities, so all of its elements are at most 1.

□

A.3. Proof of Proposition 10

Proof Let $\epsilon = \|C\|$. From Definition 8, we have that

$$\begin{aligned}
\mathcal{V}(q_1, \ldots, q_T; d, \bar{C}) &\triangleq d^\top B(q_T) \cdots B(q_1)\mathbf{1}\\
&= d^\top p_T
\end{aligned}$$

Theorem 6 in (Hayes, 2006) implies that the entries of the marginal coupling probability, p_T decay with rate $(1 - \epsilon/n)^T$ for uniform random scans. Similarly, Theorem 8 of (Hayes, 2006) implies that the entries of the marginal coupling decay with rate $(1 - \epsilon/2)^{T/n}$ for systematic scans. In both cases, the statement holds by taking T to infinity. □

A.4. Proof of Theorem 12: Influence bound for binary pairwise MRFs

Our proof relies on the following technical lemma.

Lemma 15. *Consider the function $g(w,z) = 1/(1+zw)$ for $w \geq 0$ and $z \in [r,s]$ for some $s, r \geq 0$. We have*

$$|g(w,z) - g(w',z)| = \frac{|w-w'|z}{(1+zw)(1+zw')} \leq \frac{|w-w'|z^*}{(1+z^*w)(1+z^*w')} \tag{12}$$

for $z^* = \max(r, \min(s, \sqrt{1/(ww')}))$.

Proof The inequality follows from the fact that the expression (12) is increasing in z on $[0, \sqrt{1/(ww')})$ and decreasing on $(\sqrt{1/(ww')}, \infty)$. □

Proof of Theorem 12 In the notation of Lemma 15, we see that, for each i and $j \neq i$, the full conditional of X_i is given by

$$\begin{aligned}
\pi(X_i = 1 | X_{-i}) &= \frac{1}{1 + \exp\left(-2\sum_k \theta_{ik} X_k - 2\theta_i\right)} \\
&= \frac{1}{1 + \exp\left(-2\sum_{k\neq j} \theta_{ik} X_k - 2\theta_i\right)\exp(-2\theta_{ij}X_j)} \\
&= g(\exp(-2\theta_{ij}X_j), b)
\end{aligned}$$

for $b = \exp\left(-2\sum_{k\neq j}\theta_{ik}X_k - 2\theta_i\right) \in \left[\exp\left(-2\sum_{k\neq j}\theta_{ik} - 2\theta_i\right), \exp\left(2\sum_{k\neq j}\theta_{ik} - 2\theta_i\right)\right]$.

Therefore, the influence of X_j on X_i admits the bound

$$\begin{aligned}
C_{ij} &\triangleq \max_{X,Y\in B_j} \left|\pi(X_i = 1|X_{-i}) - \pi(Y_i = 1|Y_{-i})\right| \\
&= \max_{X,Y\in B_j} |g(\exp(-2\theta_{ij}X_j), b) - g(\exp(-2\theta_{ij}Y_j), b)| \\
&= \max_{X,Y\in B_j} \frac{|\exp(-2\theta_{ij}X_j) - \exp(-2\theta_{ij}Y_j)|b}{(1+b\exp(-2\theta_{ij}X_j))(1+b\exp(-2\theta_{ij}Y_j))} \\
&= \max_{X,Y\in B_j} \frac{|\exp(2\theta_{ij}) - \exp(-2\theta_{ij})|b}{(1+b\exp(2\theta_{ij}))(1+b\exp(-2\theta_{ij}))} \\
&\leq \frac{|\exp(2\theta_{ij}) - \exp(-2\theta_{ij})|b^*}{(1+b^*\exp(2\theta_{ij}))(1+b^*\exp(-2\theta_{ij}))}
\end{aligned}$$

for $b^* = \max(\exp\left(-2\sum_{k\neq j}\theta_{ik} - 2\theta_i\right), \min(\exp\left(2\sum_{k\neq j}\theta_{ik} - 2\theta_i\right), 1))$. □

A.5. Proof of Theorem 13: Influence bound for generalized Potts models

Proof of Theorem 13 For any pair of states $a, b \in \mathcal{X}$ and distinct variable indices i, j, let

$$r_{ia}(b) \triangleq \exp\left(\theta_{i,a} + \theta_{ij}\mathbb{I}[b=a] + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = a\}\right)$$

and

$$R_i(b) \triangleq \sum_{a\in\mathcal{X}} r_{ia}(b)$$

so that $\pi(X_i = a | X_j = b, X_{-\{i,j\}}) = r_{ia}(b)/R_i(b)$.

Now consider any distinct pair of states $a, b \in \mathcal{X}$, and, without loss of generality, suppose that $\theta_{i,b} \leq \theta_{i,a}$. Since $r_{ic}(a) = r_{ic}(b)$

whenever $c \notin \{a, b\}$, we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{c\in\mathcal{X}}\left|\pi(X_i = c|X_j = a, X_{-\{i,j\}}) - \pi(X_i = c|X_j = b, X_{-\{i,j\}})\right| \\
&= \frac{1}{2}\sum_{c\in\mathcal{X}}\left|\frac{r_{ic}(a)}{R_i(a)} - \frac{r_{ic}(b)}{R_i(b)}\right| = \frac{1}{2}\sum_{c\in\mathcal{X}}\left|\frac{r_{ic}(a)}{R_i(a)} - \frac{r_{ic}(a)}{R_i(b)} + \frac{r_{ic}(a)}{R_i(b)} - \frac{r_{ic}(b)}{R_i(b)}\right| \\
&\leq \frac{1}{2}\sum_{c\in\mathcal{X}}\frac{r_{ic}(a)}{R_i(a)}\left|1 - \frac{R_i(a)}{R_i(b)}\right| + \left|\frac{r_{ic}(a) - r_{ic}(b)}{R_i(b)}\right| \\
&= \frac{1}{2}\left(\left|\frac{R_i(b) - R_i(a)}{R_i(b)}\right| + \frac{|r_{ia}(a) - r_{ia}(b)|}{R_i(b)} + \frac{|r_{ib}(a) - r_{ib}(b)|}{R_i(b)}\right) \\
&= \frac{1}{2}\left(\frac{|r_{ia}(a) - r_{ia}(b) + r_{ib}(a) - r_{ib}(b)|}{R_i(b)} + \frac{|r_{ia}(a) - r_{ia}(b)|}{R_i(b)} + \frac{|r_{ib}(a) - r_{ib}(b)|}{R_i(b)}\right) \\
&= \frac{1}{2}\left(\frac{(\exp(\theta_{ij}) - 1)|r_{ia}(b) - r_{ib}(a)|}{R_i(b)} + \frac{|\exp(\theta_{ij}) - 1|(r_{ia}(b) + r_{ib}(a))}{R_i(b)}\right) \\
&= \frac{(\exp(\theta_{ij}) - 1)\max(r_{ia}(b), r_{ib}(a))}{R_i(b)} \\
&= \frac{(\exp(\theta_{ij}) - 1)\max(r_{ia}(b), r_{ib}(a))}{r_{ia}(b) + \exp(\theta_{ij})r_{ib}(a) + \sum_{c\notin\{a,b\}}\exp\Big(\theta_{i,c} + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = c\}\Big)} \\
&\leq \frac{(\exp(\theta_{ij}) - 1)\max(r_{ia}(b), r_{ib}(a))}{r_{ia}(b) + \exp(\theta_{ij})r_{ib}(a) + \sum_{c\notin\{a,b\}}\exp(\theta_{i,c})} \\
&= \frac{(\exp(\theta_{ij}) - 1)\max\left(\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = a\}\Big), \exp\Big(\theta_{i,b} + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = b\}\Big)\right)}{\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = a\}\Big) + \exp(\theta_{ij})\exp\Big(\theta_{i,b} + \sum_{k\neq j}\theta_{ik}\mathbb{I}\{X_k = b\}\Big) + \sum_{c\notin\{a,b\}}\exp(\theta_{i,c})} \\
&\leq \max\left(\frac{(\exp(\theta_{ij}) - 1)\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big)}{\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big) + \exp(\theta_{ij})\exp(\theta_{i,b}) + \sum_{c\notin\{a,b\}}\exp(\theta_{i,c})},\right. \\
&\qquad\left.\frac{(\exp(\theta_{ij}) - 1)\exp\Big(\theta_{i,b} + \sum_{k\neq j}\theta_{ik}\Big)}{\exp(\theta_{i,a}) + \exp(\theta_{ij})\exp\Big(\theta_{i,b} + \sum_{k\neq j}\theta_{ik}\Big) + \sum_{c\notin\{a,b\}}\exp(\theta_{i,c})}\right) \\
&= \frac{\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big)(\exp(\theta_{ij}) - 1)}{\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big) + \exp(\theta_{ij})\exp(\theta_{i,b}) + \sum_{c\notin\{a,b\}}\exp(\theta_{i,c})} \\
&= \frac{\exp\Big(\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big)(\exp(\theta_{ij}) - 1)}{\exp(\theta_{i,a})(\exp\Big(\sum_{k\neq j}\theta_{ik}\Big) - 1) + \exp(\theta_{i,b})(\exp(\theta_{ij}) - 1) + \sum_{c\in\mathcal{X}}\exp(\theta_{i,c})}
\end{aligned}
$$

where the final two inequalities follow from the nonnegativity of θ_{ik} for all k. Now we note that

$$
\begin{aligned}
C_{ij} &= \max_{a,b\in\mathcal{X},X_{-\{i,j\}}}\frac{1}{2}\sum_{c\in\mathcal{X}}\left|\pi(X_i = c|X_j = a, X_{-\{i,j\}}) - \pi(X_i = c|X_j = b, X_{-\{i,j\}})\right| \\
&\leq \frac{\exp\Big(\max_{a\in\mathcal{X}}\theta_{i,a} + \sum_{k\neq j}\theta_{ik}\Big)(\exp(\theta_{ij}) - 1)}{\exp(\max_{a\in\mathcal{X}}\theta_{i,a})(\exp\Big(\sum_{k\neq j}\theta_{ik}\Big) - 1) + \exp(\min_{b\in\mathcal{X}}\theta_{i,b})(\exp(\theta_{ij}) - 1) + \sum_{c\in\mathcal{X}}\exp(\theta_{i,c})}
\end{aligned}
$$

to complete the proof. □

B. Additional experiments

We provide a few experimental results that were excluded from the main body of the paper due to space limitations.

B.1. Independent replicates of evaluating and optimizing Gibbs sampler scans experiment

Figure 8 displays the results of nine independent replicates of the "Evaluating and optimizing Gibbs sampler scans" experiment of Section 5, with independently drawn unary and binary potentials.

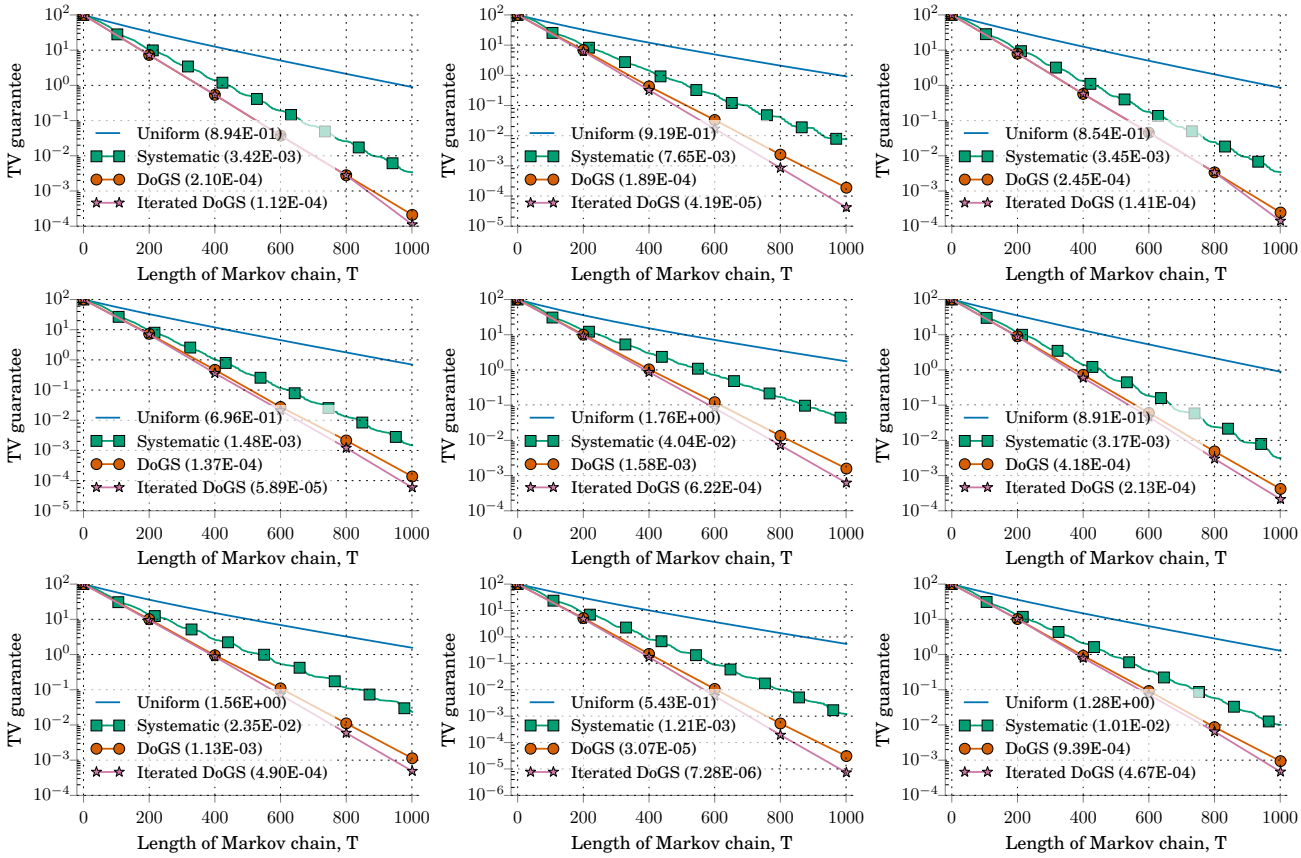

Figure 8. Independent replicates of the "Evaluating and optimizing Gibbs sampler scans" experiment of Section 5.

B.2. Independent replicates of end-to-end wall clock time performance experiment

Figure 9 repeats the timing experiment of Figure 2 providing three extra independent trials.

Figure 10 reports the estimate of a marginal expectation versus time from a sampling experiment, including the setup time for DoGS. The setting is the same as in the timing experiment in Section 5 with the exception of model size: here we simulate a 300×300 Ising model, with $90K$ variables in total. The bottom plot uses the average measured time for a single step the measured setup time for DoGS and the size of the two scan sequences ($190K$ for systematic, 16 for DoGS) to give an estimate of the speedup as a function of the number of samples we draw.

B.3. Addendum to marginal experiments

In this section we provide alternative configurations and independent runs of the marginal experiments presented in Section 5.4.

Figure 11 gives a spatial histogram of samples at different segments of the DoGS sequence. We note how the sequence starts off sampling for the target (top-left) site's extended neighborhood and slowly zeroes in on the target near the end.

Here we repeat the marginal Ising model experiments from Section 5. The set up is exactly the same, with the exception of the Ising model boundary conditions. Results are shown in Figure 14 and Figure 13.

Finally, in Figure 12, we repeat the sampling experiment of Figure 5 three times.

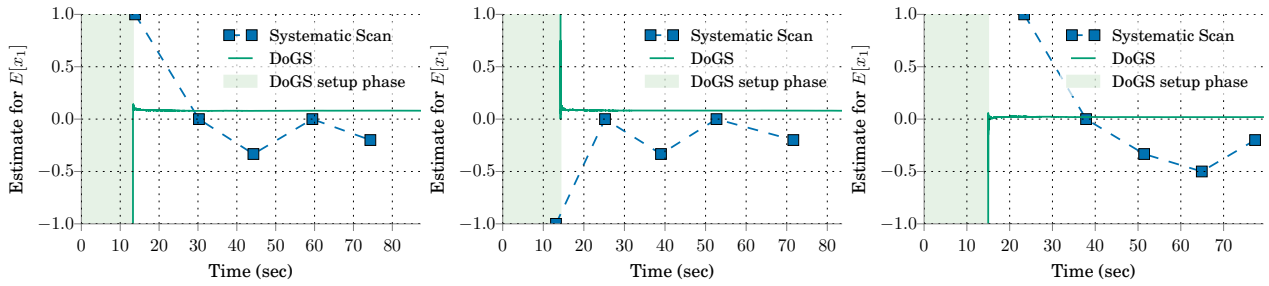

Figure 9. Estimate of $\mathbb{E}[x_1]$ versus wall-clock time for a standard row-major-order systematic scan and a DoGS optimized sequence on a $1K \times 1K$ Ising model. By symmetry $\mathbb{E}[x_1] = 0$. Three independent repetitions.

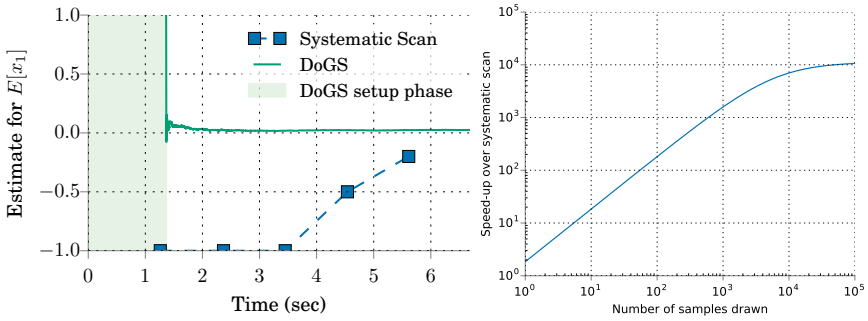

Figure 10. (top) Estimate of $\mathbb{E}[x_1]$ versus wall-clock time for a standard row-major-order systematic scan and a DoGS optimized sequence on a 300×300 Ising model. By symmetry $\mathbb{E}[x_1] = 0$. (bottom) The end-to-end speedup of DoGS over systematic scan, including setup and optimization time, as a function of the number of samples we draw.

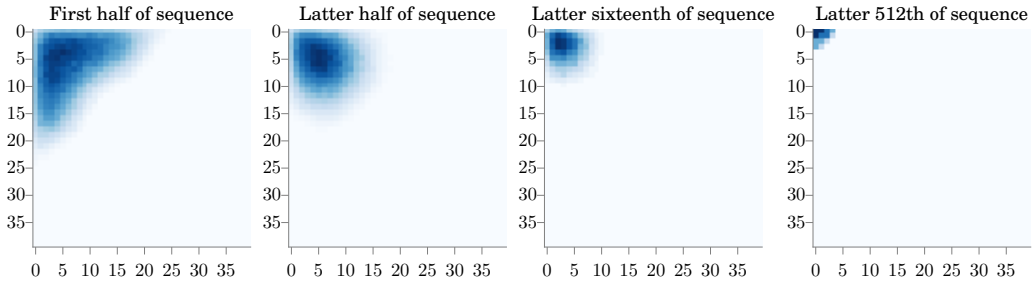

Figure 11. Sampling histogram of sequence from Algorithm 2 at different times.

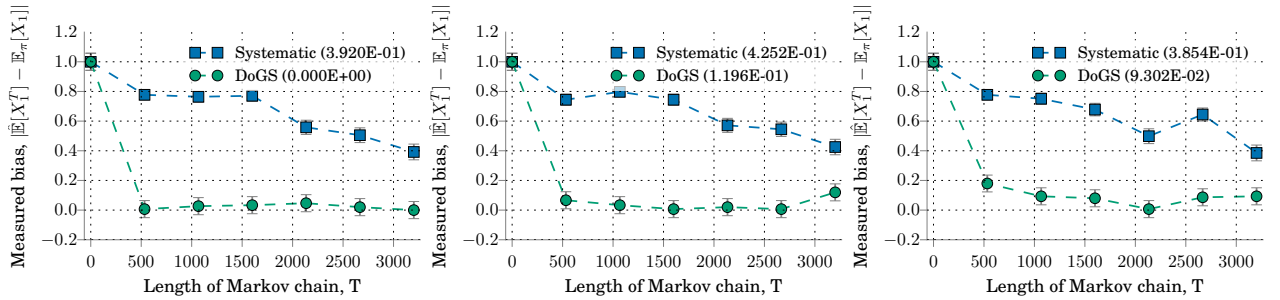

Figure 12. Three independent repetitions of the sampling experiment of Figure 5.

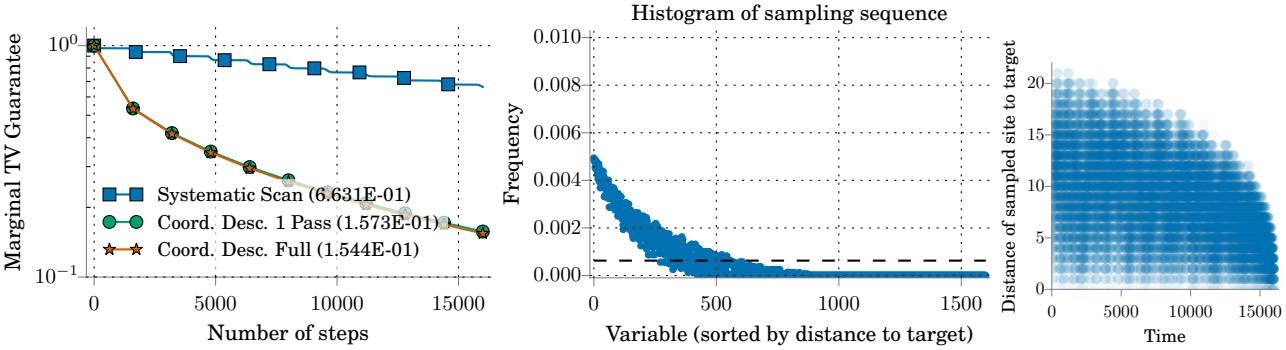

Figure 13. Deterministic scan bias comparison when targetting the top left corner variable on a 40×40 toroidal Ising model. The middle plot shows the histogram of the sequence achieved via Algorithm 2. The right plot shows the sequence's distance-time profile.

B.4. Addendum to segmentation experiments

Figure 15 provides two extra independent runs of the image segmentation experiment from Section 5.

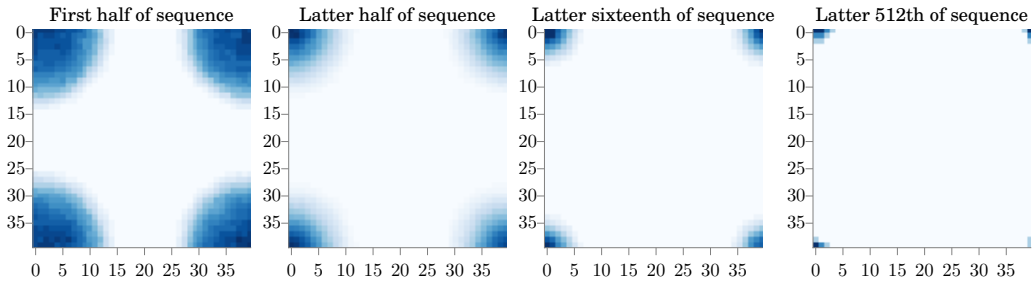

Figure 14. Sampling histogram of sequence from Algorithm 2 at different times.

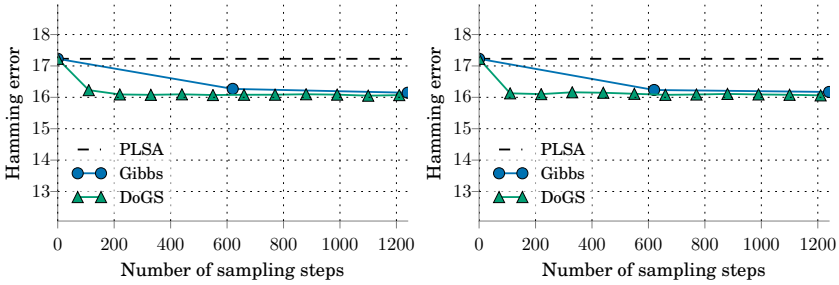

Figure 15. Average of segmentation Hamming error over 480 images for PLSA (image-based baseline) and the MRF model of (Verbeek & Triggs, 2007) trained by systematic scan over the whole image, and DoGS optimized for a 12×8 center patch.