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 discrete 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 & 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 is the *scan*, the order in which variables are sampled (He et al., 2016). 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 were 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 for discrete target distributions based on Dobrushin's notion of variable influence (Dobrushin & Shlosman, 1985). We show that for a given subset of variables, this *Dobrushin variation* (DV) 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.

The improvements achieved by DoGS are driven by an inputted matrix, $\bar{C}$, of pairwise variable influence bounds discussed in more detail in Section 3. While DoGS can be used with any discrete distribution, it was designed for targets with total influence $\|\bar{C}\| < 1$, measured in any matrix norm. This criterion is known to hold for a variety of distributions, including Ising models with sufficiently high temperatures, hard-core lattice gas models, random graph colorings (Hayes, 2006), and classes of weighted constraint satisfaction problems (Feng et al., 2017). Moreover, as we will see in Section 4.1, suitable variable influence bounds are readily available for pairwise and binary Markov random fields. These user-friendly bounds give rise to total

[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>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

influence $\|\bar{C}\| < 1$ in all of our experiments and thereby enable improvements in both inferential speed and accuracy over standard scans.

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 (weighted) 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 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 $\mathrm{diag}(v)$ for a square diagonal matrix with v on the diagonal and $\odot$ for element-wise multiplication. 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 matrix C. We use the shorthand $[p] \triangleq \{1, \dots, p\}$.

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 $X^0, X^1, \dots, X^T \in \mathcal{X}^p$ from tractable distributions such that expectations over X^T are close to expectations under π.

2.1. Gibbs sampling

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 $q_1, \dots, q_T$ determining the probability of resampling each variable on each round of Gibbs sampling. Typically one selects between the uniform random scan, $q_t = (1/p, \dots, 1/p)$ 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 $\mu$; single-variable
       conditionals of target distribution, $\pi(\cdot|X_{-i})$
  Sample from starting distribution: $X^0 \sim \mu$
  for $t$ in $1, 2, \dots, T$ do
    Sample variable index to update using scan: $i_t \sim q_t$
    Sample $X_{i_t}^t \sim \pi(\cdot|X_{-i_t}^{t-1})$ from its conditional
    Copy remaining variables: $X_{-i_t}^t = X_{-i_t}^{t-1}$
  end for
output  Sample sequence $(X^t)_{t=0}^T$
```

2.2. Total variation

Let π_t represent the distribution of the t-th step, X^t, of a Gibbs sampler. The quality of a T-step Gibbs sampler and its scan is typically measured in terms of total variation (TV) 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,*

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

We view TV as providing a bound on the bias of a large class of Gibbs sampler expectations; note, however, that TV does not control the variance of these expectations.

2.3. Marginal and weighted 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 segmenting an image into its component parts, a particular region, 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 2 (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:*

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

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

Definition 3 (d-bounded differences)**.** *We say $f : \mathcal{X}^p \to \mathbb{R}$ has* d-bounded differences *for $d \in \mathbb{R}^d$ if, for all $X, Y \in \mathcal{X}^p$,*

$$|f(X) - f(Y)| \leq \sum_{i=1}^{p} d_i \mathbb{I}[X_i \neq Y_i].$$

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 tailored to d-bounded difference functions.

Definition 4 (d-weighted total variation)**.** *The* d-weighted total variation *between probability measures μ and ν is the maximum difference in expectations across d-bounded difference functions:*

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

3. Measuring scan quality with Dobrushin variation

Since the direct computation of total variation measures is typically prohibitive, we will define an efficiently computable upper bound on the weighted total variation of Definition 4. Our construction is inspired by the Gibbs sampler convergence analysis of Dobrushin & Shlosman (1985).

The first step in Dobrushin's approach is to control total variation in terms of coupled random vectors, $(X_t, Y_t)_{t=0}^T$, where X^t has the distribution, π_t, of the t-th step of the Gibbs sampler and Y^t follows the target distribution π. For any such coupling, we can define the marginal coupling probability $p_{t,i} \triangleq \mathbb{P}(X_i^t \neq Y_i^t)$. The following lemma, a generalization of results in (Dobrushin & Shlosman, 1985; Hayes, 2006), shows that weighted total variation is controlled by these marginal coupling probabilities. The proof is given in Appendix A.1, and similar arguments can be found in Rebeschini & van Handel (2014).

Lemma 5 (Marginal coupling controls weighted TV)**.** *For any joint distribution (X, Y) such that $X \sim \mu$ and $Y \sim \nu$ for probability measures μ and ν on $\mathcal{X}^p$ and any nonnegative weight vector $d \in \mathbb{R}^p$,*

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

Dobrushin's second step is to control the marginal coupling probabilities p_t in terms of *influence,* a measure of how much a change in variable j affects the conditional distribution of variable i.

Definition 6 (Dobrushin influence matrix)**.** *The* Dobrushin influence *of variable j on variable 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 $l \neq j$.

This influence matrix is at the heart of our efficiently computable measure of scan quality, *Dobrushin variation.*

Definition 7 (Dobrushin variation)**.** *For any nonnegative weight vector $d \in \mathbb{R}^p$ and entrywise upper bound $\bar{C}$ on the Dobrushin influence* (1)*, we define the* Dobrushin variation *of a scan $(q_t)_{t=1}^T$ as*

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

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

Theorem 8 shows that Dobrushin variation dominates weighted TV and thereby provides target- and scan-specific guarantees on the weighted TV quality of a Gibbs sampler. The proof in Appendix A.2 rests on the fact that, for each t, $b_t \triangleq B(q_t) \cdots B(q_1)\mathbf{1}$ provides an elementwise upper bound on the vector of marginal coupling probabilities, p_t.

Theorem 8 (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 $d \in \mathbb{R}^p$ and entrywise upper bound $\bar{C}$ on the Dobrushin influence* (1)*,*

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

4. Improving scan quality with DoGS

We next present an efficient algorithm for improving the quality of any Gibbs sampler scan by minimizing Dobrushin variation. 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. Since Dobrushin variation is linear in each q_t, each coordinate optimization (in the absence of ties) selects a degenerate distribution, a single coordinate, yielding a fully deterministic scan. If $m \leq p$ is a bound on the size of the Markov blanket of each variable, then our forward-backward algorithm runs in time $O(\|d\|_0 + \min(m \log p + m^2, p)T)$ with $O(p + T)$ storage for deterministic input scans. The $T(m \log p + m^2)$ term arises from maintaining the derivative vector, w, in an efficient sorting structure, like a max-heap.

A user can initialize DoGS with any baseline scan, including a 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

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 $\epsilon$.

  // Forward: Precompute coupling bounds of Section 3,
  // $b_t = B(q_t)\cdots B(q_1)\mathbf{1} = B(q_t)b_{t-1}$ with $b_0 = \mathbf{1}$.
  // Only store $b = b_T$ and sequence of changes $(\Delta_t^b)_{t=0}^{T-1}$.
  // Also precompute Dobrushin variation $\mathcal{V} = d^\top b_T$
  // and derivatives $w = \partial\mathcal{V}/\partial q_T = -d \odot (I - \bar{C})b_T$.
  $b \leftarrow \mathbf{1}$, $\mathcal{V} \leftarrow d^\top b$, $w \leftarrow -d \odot (I - \bar{C})b$
  for $t$ in $1, 2, \ldots T$ do
    $\Delta_{t-1}^b \leftarrow \mathrm{diag}\,(q_t)(I - \bar{C})b$
    $b \leftarrow b - \Delta_{t-1}^b$                          // $b_t = b_{t-1} - \Delta_{t-1}^b$
    $\mathcal{V} \leftarrow \mathcal{V} - d^\top \Delta_{t-1}^b$                      // $\mathcal{V} = d^\top b_t$
    $w \leftarrow w + d \odot (I - \bar{C})\Delta_{t-1}^b$   // $w = -d \odot (I - \bar{C})b_t$
  end for

  // Backward: Optimize scan one step, $q_t^*$, at a time.
  for $t$ in $T, T-1, \ldots, 1$ do
    If $\mathcal{V} \leq \epsilon$, then $q_t^* \leftarrow q_t$; break          // early stopping
    $b \leftarrow b + \Delta_{t-1}^b$                          // $b_{t-1} = b_t + \Delta_{t-1}^b$
    // Update $w = \partial\mathcal{V}/\partial q_t = -d_t \odot (I - \bar{C})b_{t-1}$
    // for $d_t^\top \triangleq d^\top B(q_T^*)\cdots B(q_{t+1}^*)$ and $d_T^\top \triangleq d^\top$
    $w \leftarrow w - d \odot (I - \bar{C})\Delta_{t-1}^b$
    // Pick probability vector $q_t^*$ minimizing $d_t^\top B(q_t)b_{t-1}$
    $q_t^* \leftarrow e_{\mathrm{argmin}_i\, w_i}$
    $\mathcal{V} \leftarrow \mathcal{V} + d^\top \mathrm{diag}(q_t^* - q_t)b$             // $\mathcal{V} = d_{t-1}^\top b_{t-1}$
    $\Delta^{d^\top} \leftarrow d^\top \mathrm{diag}(q_t^*)(I - \bar{C})$
    $d^\top \leftarrow d^\top - \Delta^{d^\top}$                       // $d_{t-1}^\top = d_t^\top B(q_t^*)$
    $w \leftarrow w + \Delta^d \odot (I - \bar{C})b$ // $w = -d_{t-1} \odot (I - \bar{C})b_{t-1}$
  end for
output  Optimized scan $(q_\tau)_{\tau=1}^{t-1}, (q_\tau^*)_{\tau=t}^T$
```

be d-ergodic (i.e., $\|\pi_T - \pi\|_{d,\mathrm{TV}} \to 0$ 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 8—goes to 0 under these conditions and standard scans. The proof relies on arguments in (Hayes, 2006) and is outlined in Appendix A.3.

Proposition 9. *Suppose that $\bar{C}$ is an entrywise upper bound on the Dobrushin influence matrix (1) and that $(q_t)_{t=1}^T$ 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. That is,*

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

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 useful influence bounds are particularly straightforward to compute for any pairwise Markov random field (MRF) target,

$$\pi(X) \propto \exp(\textstyle\sum_{i,j}\sum_{a,b\in\mathcal{X}} \theta_{ab}^{ij}\,\mathbb{I}[X_i = a, X_j = b]). \quad (2)$$

Theorem 10 (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 (2) satisfies*

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

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

$$\pi(X) \propto \exp(\textstyle\sum_{i\neq j}\theta_{ij}X_iX_j + \sum_i \theta_i X_i), \quad (3)$$

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 11 (Binary pairwise influence)**.** *The influence (1) of the target π in (3) satisfies*

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

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 11 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 $\mathcal{B}$.

So far, we have focused on bounding influence in pairwise MRFs, as these bounds are most relevant to our experiments; indeed, in Section 5, we will use DoGS in conjunction with the bounds of Theorems 10 and 11 to improve scan quality for a variety of inferential tasks. However, user-friendly bounds are also available for non-pairwise MRFs (note that any discrete distribution can be represented as an MRF with parameters in the extended reals), and we include a simple extension of Theorem 11 that applies to binary MRFs with higher-order interactions. Its proof is in Appendix A.5

Theorem 12 (Binary higher-order influence)**.** *The target*

$$\pi(X) \propto \exp(\textstyle\sum_{S\in\mathcal{S}}\theta_S\prod_{k\in S}X_k + \sum_i \theta_i X_i),$$

for $X \in \{-1, 1\}^d$ and $\mathcal{S}$ a set of non-singleton subsets of $[p]$, has influence (1) satisfying

$$C_{ij} \leq \frac{|\exp\left(2\sum_{S\in\mathcal{S}:i,j\in S}|\theta_S|\right) - \exp\left(-2\sum_{S\in\mathcal{S}:i,j\in S}|\theta_S|\right)|\, b^*}{(1+b^*)^2}$$

for $b^* = \max(\exp(-2\sum_{S\in\mathcal{S}:i\in S, j\notin S}|\theta_S| - 2\theta_i), \min(\exp(2\sum_{S\in\mathcal{S}:i\in S, j\notin S}|\theta_S| - 2\theta_i), 1))$.

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 adopt the model parameterization of (3) (with no additional temperature parameter) and use Theorem 11 to produce the Dobrushin influence bound $\bar{C}$. On all ensuing plots, the numbers in the legend state 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.

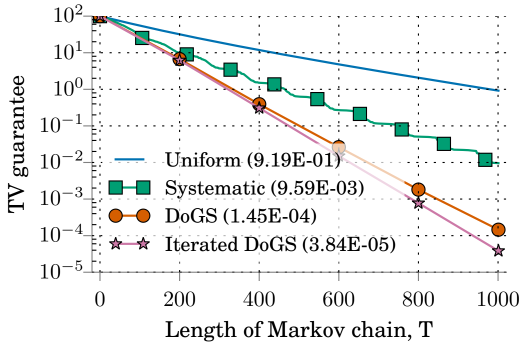

Figure 1. TV guarantees provided Dobrushin variation for various Gibbs sampler scans on a 10×10 non-toroidal Ising model with random parameters (see Section 5.1). DoGS is initialized with the systematic scan.

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 (3), 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 8) 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 $d = \mathbf{1}$ and 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 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 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 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.

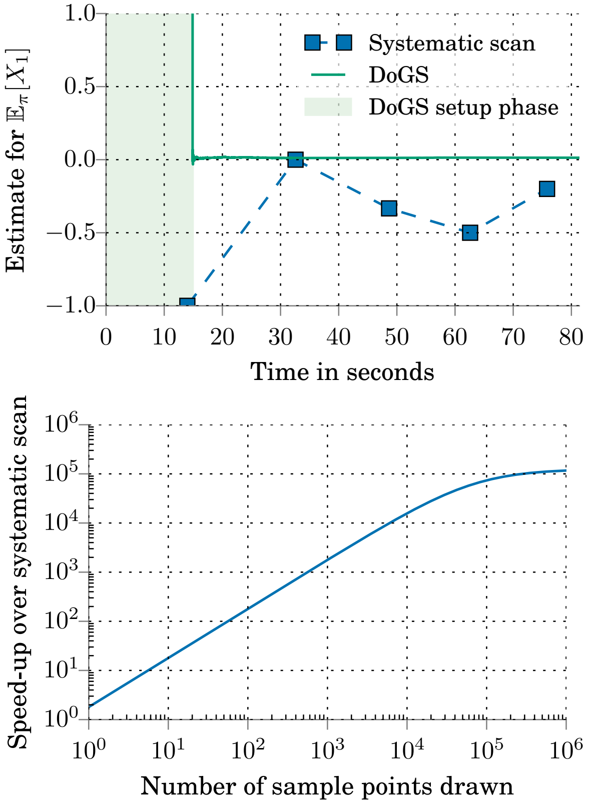

Figure 2. (top) Estimate of target, $\mathbb{E}_\pi[X_1]$, versus wall-clock time for a standard row-major-order systematic scan and a DoGS optimized sequence on an Ising model with 1 million variables (see Section 5.2). By symmetry $\mathbb{E}_\pi[X_1] = 0$. (bottom) The end-to-end speedup of DoGS over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

5.2. End-to-end wall clock time performance

In this experiment, we demonstrate that using DoGS to optimize a scan can result in dramatic inferential speed-ups. This effect is particularly pronounced for targets with a large number of variables and in settings that require repeated sampling from a low-bias Gibbs sampler. The setting is the exactly same as in the previous experiment, with the exception of model size: here we simulate a $10^3 \times 10^3$ Ising model, with 1 million variables in total. We target a single marginal X_1 with $d = e_1$ and take a systematic scan of length $T = 2 \times 10^6$ as our input scan. After measuring the Dobrushin variation ϵ of the systematic scan, we use an efficient length-doubling scheme to select a DoGS scan: (0) initialize $\tilde{T} = 2$; (1) run Algorithm 2 with the first $\tilde{T}$ steps of the systematic scan as input; (2) if the resulting DoGS scan has Dobrushin variation less than ϵ, we keep it; otherwise we double $\tilde{T}$ and return to step (1). The resulting DoGS scan has length $\tilde{T} = 16$.

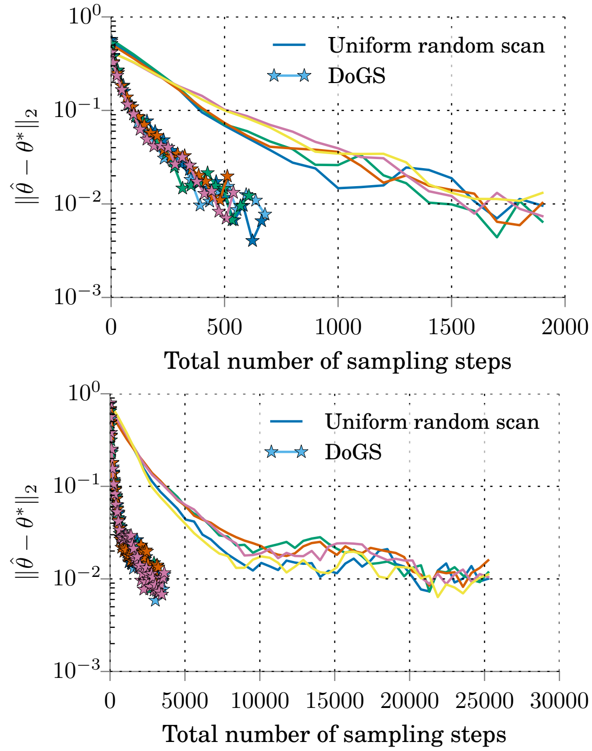

Figure 3. Comparison of parameter estimation error in MCMC maximum likelihood estimation of the 3×3 (top) and a 4×4 (bottom) Ising models of Domke (2015). Each MCMC gradient estimate is obtained either from the uniform random scan suggested by Domke or from DoGS initialized with the uniform random scan,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

We repeatedly draw independent sample points from either the length T systematic scan Gibbs sampler or the length $\tilde{T}$ DoGS scan Gibbs sampler. Figure 2 evaluates the bias of the resulting Monte Carlo estimates of $\mathbb{E}_\pi[X_1]$ as a function of time, including the 15s of setup time for DoGS on this 1 million variable model. In 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 to give an estimate of the speedup as a function of the number of sample points drawn. Additional timing experiments are deferred to Appendix B.2.

5.3. Accelerated MCMC maximum likelihood estimation

We next illustrate how DoGS can be used to accelerate MCMC maximum likelihood estimation, while providing guarantees on parameter estimation quality. We replicate

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

the Ising model maximum likelihood estimation experiment of (Domke, 2015, Sec. 6) and show how we can provide the same level of accuracy faster. Our aim is to learn the parameters of binary MRFs based on training samples with independent Rademacher entries. On each step of MCMC-MLE, Domke uses Gibbs sampling with a uniform random scan to produce an estimate of the gradient of the log likelihood. Our DoGS variant employs Algorithm 2 with $d = \mathbf{1}$, early stopping parameter $\epsilon = 0.01$, and a Dobrushin influence bound constructed from the latest parameter estimate $\hat{\theta}$ using Theorem 11. We set the number of gradient steps, MC steps per gradient, and independent runs of Gibbs sampling to the suggested values in (Domke, 2015). After each gradient update, we record the distance between the optimal and estimated parameters. Figure 3 displays the estimation error of five independent replicates of this experiment using each of two scans (uniform or DoGS) for two models (a 3×3 and a 4×4 Ising model). 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 target-dependent guarantees. We use a 40×40 non-toroidal Ising model and set our feature to be the top left variable with $d = e_1$. 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 additional experiments on a toroidal Ising model in Appendix B.3.

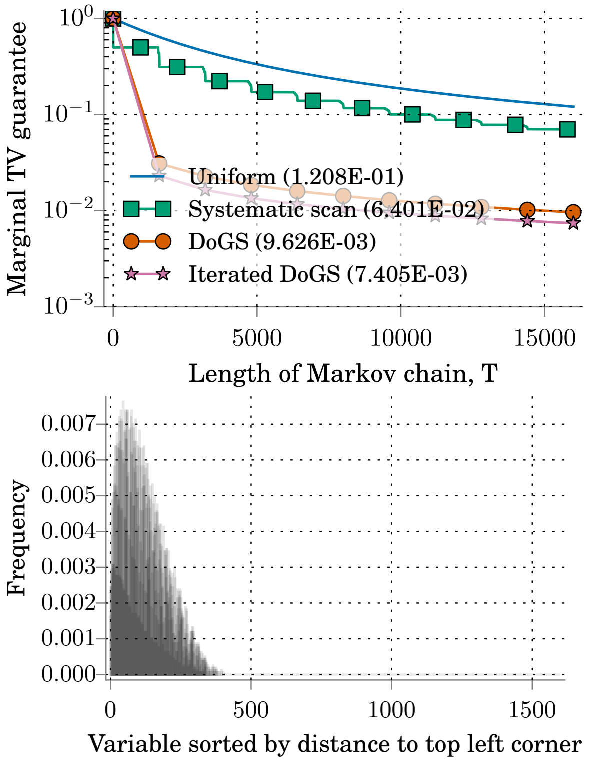

Figure 4. (top) Marginal TV guarantees provided by Dobrushin variation for various Gibbs sampler scans when targeting the top left corner variable on a 40×40 non-toroidal Ising model with $\theta_{ij} \approx 1/3.915$ (see Section 5.4). DoGS is initialized with the systematic scan. (bottom) Frequency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

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 each patch, and induces the posterior distribution

$$\pi(X|y;k) \propto \exp\big(\textstyle\sum_{(i,j)\text{ spatial neighbors}} \sigma \mathbb{I}\{X_i = X_j\} \quad (4)$$
$$+ \textstyle\sum_i \log(\sum_{a \in \mathcal{X}} \theta_{k,a} \beta_{a,y_i} \mathbb{I}\{X_i = a\})\big),$$

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 (4) 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/

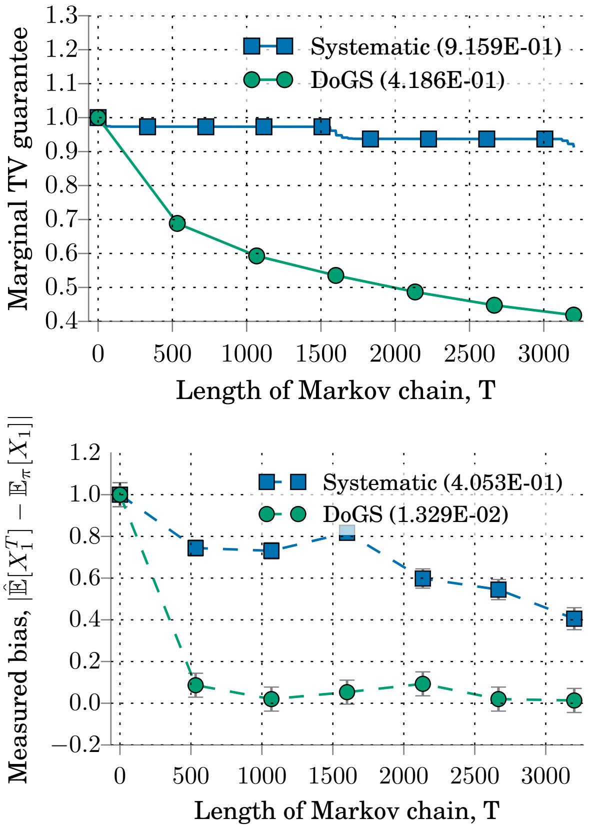

Figure 5. (top) Marginal TV guarantees provided by Dobrushin variation for systematic scan and DoGS initialized with systematic scan when targeting the top left corner variable on a 40×40 toroidal Ising model with $\theta_{ij} = 0.25$ (see Section 5.4). (bottom) Measured bias and standard errors from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

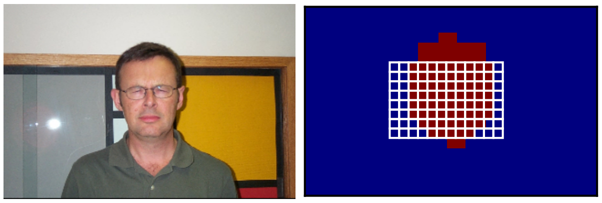

Figure 6. (left) Example test image from MSRC dataset. (right) Segmentation produced by DoGS Markov field aspect model targeting the center region outlined in white (see Section 5.5).

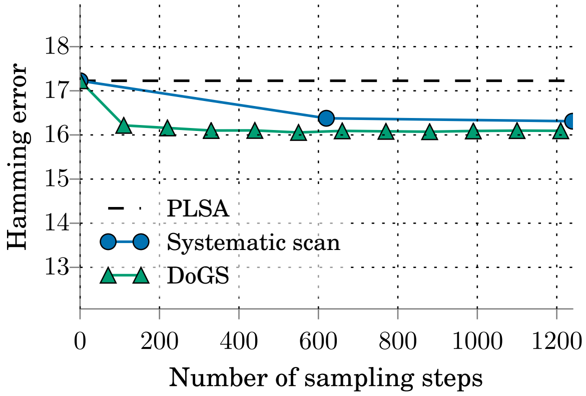

Figure 7. Average test image segmentation error under the Markov field aspect model of Section 5.5. PLSA represents the maximum a posteriori patch labeling under the MFAM (4) with $\sigma = 0$. Errors are averaged over 20 MSRC test sets of 24 images.

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 (the {0,1} entries of d indicate whether a patch is in the marginal rectangle highlighted in 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 10, 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 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

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.

Feng, Weiming, Sun, Yuxin, and Yin, Yitong. What can be sampled locally? *arXiv preprint arXiv:1702.00142*, 2017.

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.

He, Bryan D, De Sa, Christopher M, Mitliagkas, Ioannis, and Ré, Christopher. Scan order in gibbs sampling: Models in which it matters and bounds on how much. In *Advances in Neural Information Processing Systems*, pp. 1–9, 2016.

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.

Rebeschini, Patrick and van Handel, Ramon. Comparison theorems for gibbs measures. *Journal of Statistical Physics*, 157(2):234–281, 2014.

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.