Annular Augmentation Sampling

Francois Fagan
Columbia University

Jalaj Bhandari
Columbia University

John P. Cunningham
Columbia University

Abstract

The exponentially large sample space of general binary probabilistic models renders intractable standard operations such as exact marginalization, inference, and normalization. Typically, researchers deal with these distributions via deterministic approximations, the class of belief propagation methods being a prominent example. Comparatively, Markov Chain Monte Carlo methods have been significantly less used in this domain. In this work, we introduce an auxiliary variable MCMC scheme that samples from an annular augmented space, translating to a great circle path around the hypercube of the binary sample space. This annular augmentation sampler explores the sample space more effectively than coordinate-wise samplers and has no tunable parameters, leading to substantial performance gains in estimating quantities of interest in large binary models. We extend the method to incorporate into the sampler any existing mean-field approximation (such as from belief propagation), leading to further performance improvements. Empirically, we consider a range of large Ising models and an application to risk factors for heart disease.

1 INTRODUCTION

Binary probabilistic models are a fundamental and widely used framework for encoding dependence between discrete variables, with applications from physics (Wang and Landau, 2001) and computer vision (Nowozin and Lampert, 2011) to natural language processing (Johnson et al., 2007). The exponential nature of these distributions — a d-dimensional binary model has sample space of size 2^d — renders intractable key operations like exact marginalization and exact inference. Indeed these operations are formally hard: calculating the partition function is in general #P-complete (Chandrasekaran et al., 2008). This intractability has prompted extensive research into approximation techniques, with deterministic approximation methods such as belief propagation (Murphy et al., 1999; Wainwright and Jordan, 2008) and its variants (Murphy, 2012) enjoying substantial impact. Weighted model counting is another promising approach that has received increasing attention (Chavira and Darwiche, 2008; Chakraborty et al., 2014). However, Markov Chain Monte Carlo (MCMC) methods have been relatively unexplored, and in many settings basic Metropolis Hastings (MH) is still the default choice of sampler for binary probabilistic models (Zhang et al., 2012).

This scarcity of MCMC methods contrasts with the literature on continuous probabilistic models, where MCMC methods like Hamiltonian Monte Carlo (HMC), (Neal, 2011) and slice sampling (Neal, 2003) have enjoyed substantial success. Whether explicitly or implicitly stated, the key idea for many of these state-of-the-art continuous samplers is to utilize a two-operator construction: at each iteration of the algorithm, first a subset of the full sample space is chosen, and second a new point is sampled from that subspace. Consider HMC, which samples a point from the Hamiltonian flow (the subset) induced by a particular sample of the momentum variable. This two-operator construction is enabled by an auxiliary variable augmentation (e.g. the momentum variable in HMC) that separates each MCMC operator. Recently this two-operator concept for HMC has been modified to sample binary distributions (Zhang et al., 2012; Pakman and Paninski, 2013), and its improved performance over MH and Gibbs sampling has been demonstrated, giving promise to the notion of auxiliary augmentations for binary samplers.

Another continuous example of this two-operator formulation, which inspires our present work, is Elliptical Slice Sampling (ESS), (Murray et al., 2010): auxiliary variables are introduced into a latent Gaussian model such that an elliptical contour (the subset) is defined by the first sampling operator, and then the second

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

operator draws the next sample point from that ellipse.

Here we leverage this central idea of auxiliary augmentation to create a two-operator MCMC sampler for binary distributions. At each iteration, we first choose a subset of $2d$ points ($\ll 2^d$, the size of the sample space) defined over an annulus in the auxiliary variable space (or alternatively, a great circle around the hypercube corresponding to the sample space), and in the second operator we sample over that annulus. Critically, this augmentation strategy leads to an overrelaxed sampler, allowing long range exploration of the sample space (flipping multiple dimensions of the binary vector), unlike more basic coordinate-by-coordinate MH strategies, which require geometrically many iterations to flip many dimensions of the binary vector. We also extend this auxiliary formulation to exploit available deterministic approximations to improve the quality of the sampler. We evaluate our sampler on a synthetic problem (2D Ising models) and a real world problem of modeling heart disease risk factors; which significantly outperforms MH and HMC techniques for binary distributions. Overall, our contributions include:

- a novel annular augmentation that maps a continuous path in the auxiliary space to the discrete sample space of binary distributions (Section 2.1);
- an extension of this augmentation to incorporate deterministic posterior approximations such as mean-field belief propagation (Section 2.1);
- a Rao-Blackwellized Gibbs sampler that exploits this annular augmentation to achieve the desired MCMC sampler, a sampler which is both free of tuning parameters and straightforward to implement with generic code (Section 2.3);
- and experimental results demonstrating the performance improvements achieved by annular augmentation sampling (Section 3).

We conclude by discussing some extensions to annular augmentation, and future directions for the class of augmented binary MCMC techniques (Section 4).

2 ANNULAR AUGMENTATION SAMPLING

We are interested in sampling from a probability distribution $p(\mathbf{s})$ defined over d-dimensional binary vectors $\mathbf{s} \in \{-1, +1\}^d$. The density $p(\mathbf{s})$ is given in terms of a function $f(\mathbf{s})$ as

$$p(\mathbf{s}) = \frac{1}{Z} f(\mathbf{s}),$$

where Z is the normalizing constant. As in other works using augmentation strategies (Pakman and Paninski, 2013; Murray et al., 2010), we rewrite $\mathbf{s}$ as a deterministic function $\mathbf{s} = h(\theta, \mathbf{t})$ of auxiliary random variables $(\theta, \mathbf{t})$ (to be defined), where both the function and the distribution of the auxiliary variables are chosen to preserve the measure $p(\mathbf{s})$ up to a normalizing constant. It is perhaps most straightforward to take this step by partitioning p into the product of a tractable distribution $\hat{p}$, which will implicitly characterize the deterministic function, and the remainder $\mathcal{L}$:

$$p(\mathbf{s}) \propto \mathcal{L}(\mathbf{s})\hat{p}\,(\mathbf{s})\,, \qquad \mathbf{s} = h(\theta, \mathbf{t}),$$

where $\mathcal{L}(\mathbf{s}) = f(\mathbf{s})/\hat{p}(\mathbf{s})$ can be thought of as a likelihood with respect to the prior $\hat{p}$. We must choose $\hat{p}$ such that it is tractable to find a $\hat{p}$-preserving function h from the auxiliary space $(\theta, \mathbf{t})$ to $\mathbf{s}$. Furthermore, we would like $\hat{p}$ to capture the structure of p (else samples will be wasted, similar to a bad importance sampler). We now demonstrate how to do so with any fully factorized $\hat{p}$ and annular auxiliary variables.

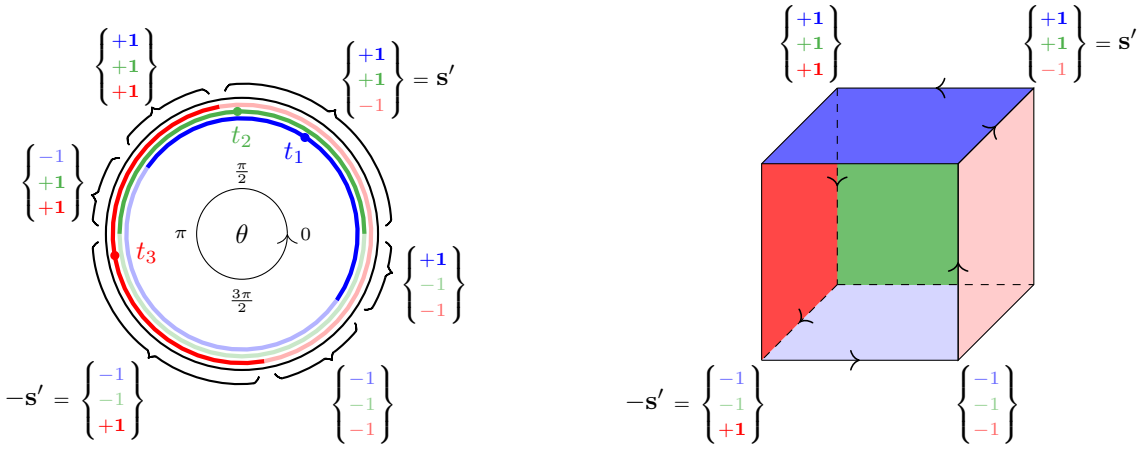
2.1 Annular Augmentations

Let us first consider the choice of a uniform prior: $\hat{p}(\mathbf{s}) = \prod_{i=1}^{d} \hat{p}(s_i)$ with $\hat{p}(s_i = +1) = \hat{p}(s_i = -1) = 1/2$. We can now replace $\mathbf{s}$ with auxiliary variables $(\theta, \mathbf{t})$ as

$$\begin{aligned} \theta &\sim unif(0, 2\pi) \\ t_i &\sim unif(0, 2\pi) \\ s_i &= h(\theta, t_i) = \text{sgn}[\cos(t_i - \theta)], \end{aligned} \tag{1}$$

where $\mathbf{t} = (t_1, ..., t_d) \in [0, 2\pi]^d$. This choice of distribution on $(\theta, \mathbf{t})$ and the map $\mathbf{s} = h(\theta, \mathbf{t})$ maintains the distribution $\hat{p}(\mathbf{s})$, but induces conditional dependence given the auxiliary variables. Critical to this work is the observation that, conditioned on a value of $\mathbf{t}$, the sample $\mathbf{s}$ can take on $2d$ possible values (*not* 2^d); which value $\mathbf{s}$ takes is determined by θ. Thus $\mathbf{t}$ defines a subset of the sample space over which we can sample $\mathbf{s}$ using θ.

Geometrically, this fact can be seen in two ways. First, Figure 1a shows each t_i as a red, green, or blue point, where the interval $t_i \pm \frac{\pi}{2}$ is shown as a similarly colored bar; this bar defines the threshold for flipping the ith coordinate of $\mathbf{s}$ (i.e., the threshold of the sgn function in Equation 1). As θ traverses $0 \to 2\pi$, $2d$ bit flips alter $\mathbf{s}'$ to $-\mathbf{s}'$ and back. Alternatively, a second view of this geometric structure is to consider a *great circle* path around the d-dimensional hypercube, as shown in Figure 1b. Owing to this fundamental ring structure, we call this auxiliary variable scheme *annular augmentation*. We use this augmentation to create an MCMC sampler in Section 2.2.



(a) Annular augmentation

(b) Path of annulus on hypercube

Figure 1: Diagram of the annuluar augmentation. (a) The auxiliary annulus. Each threshold t_i defines a semicircle where $s_i = +1$, which is indicated by a darker shade. The value of $\boldsymbol{s}$ changes as θ moves around the annulus. (b) The implied path of the annulus on the binary hypercube. Hypercube faces are colored to match the annulus.

Before introducing the sampler, we introduce an important extension. In some settings we may have access to a mean-field posterior approximation $\hat{p}$ (e.g., from loopy belief propagation), which should in general serve us better than the uniform $\hat{p}$ just introduced: a posterior approximation should "flatten" the likelihood $\mathcal{L}(\mathbf{s}) = f(\mathbf{s})/\hat{p}(\mathbf{s})$ in the spirit of (Wang and Landau, 2001) and (Fagan et al., 2016), thereby improving the efficiency of our sampler. Our annuluar augmentation generalizes nicely to this setting, by altering Eqn. 1 to:

$$s_i = \text{sgn}[\cos(t_i - \theta) - \cos(\pi\hat{p}(s_i = 1))]. \tag{2}$$

Figure 2a demonstrates that the proportion of the annulus where $s_i = 1$ becomes:

$$\frac{(t_i + \pi\hat{p}(s_i = 1)) - (t_i - \pi\hat{p}(s_i = 1))}{2\pi} = \hat{p}(s_i = 1),$$

resulting in the implied prior measure on $\mathbf{s}$ being precisely the mean-field approximation $\hat{p}$. Indeed, when $\hat{p}(s_i = 1) = 1/2$, Equation 2 reverts back to Equation 1, so the augmentations are consistent. Geometrically this change to Equation 1 induces stretched thresholds around the annulus, as shown in Figure 2b. We use the stretched annulus to good effect in Section 3.

2.2 Generic Sampler

As outlined in Section 1, to sample from $p(\mathbf{s})$ we define two MCMC operators to sample our auxiliary variables $(\theta, \mathbf{t})$, and then we read out the corresponding values of $\mathbf{s} = h(\theta, \mathbf{t})$ as given by Equation 1 or, more generally, Equation 2. The posterior over the auxiliary variables $(\theta, \mathbf{t})$ is:

$$p(\theta, \mathbf{t}) \propto \mathcal{L}(h(\theta, \mathbf{t})), \tag{3}$$

where the uniform $\frac{1}{2\pi}$ prior for θ and $\mathbf{t}$ have been absorbed into the constant of proportionality. In Operator 1, we use $\mathbf{t}$ to define a subspace, and in Operator 2 we sample over θ:

- **Operator 1:** Sample $\mathbf{t}$ subject to the constraint $\mathbf{s} = h(\theta, \mathbf{t})$ is fixed. The simplest such operator uniformly samples t_i from the domain where s_i is fixed:

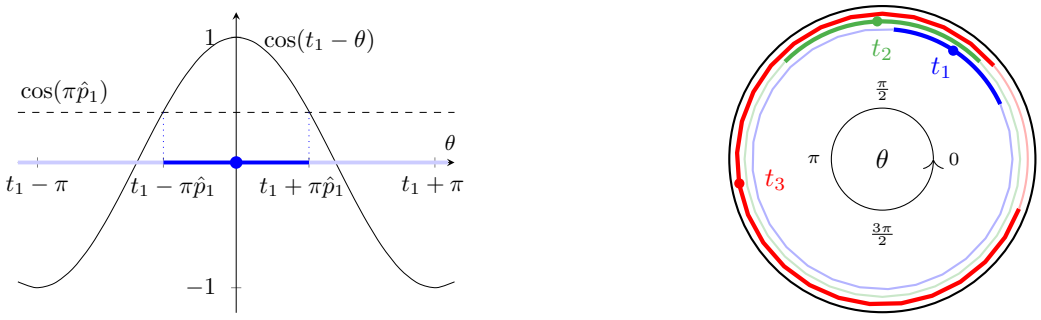
$$t_i \sim unif(\theta + (1 - s_i)\pi/2 - \pi\hat{p}(s_i), \\ \theta + (1 - s_i)\pi/2 + \pi\hat{p}(s_i)).$$

- **Operator 2**: Sample θ using an MCMC transition operator (e.g. Gibbs sampling).

First note that, in Operator 1, the density of the proposed point is equal to that of the current point since $\mathcal{L}(h(\theta, \mathbf{t}))$ is unchanged. Thus the sample from Operator 1 is accepted with probability 1 and is a valid MCMC step (which follows from evaluating the standard MH acceptance probability). Second, Operator 2 is also a valid MCMC transition by definition. Together, this two-operator sampling approach leaves the target distribution in Equation 3 invariant and is ergodic, and thus the Markov chain will converge to the stationary distribution p.

2.3 Operator 2: Rao-Blackwellized Gibbs

In the previous section we proposed a generic sampler for the annular augmentation; it now remains to specify Operator 2 to sample θ conditioned on $\mathbf{t}$. Though a number of choices are available, a particularly effective



(a) Interval where $s_i = 1$

(b) Stretched thresholds

Figure 2: Augmentation with mean-field prior $\hat{p}$. (a) Equation 2 in graphical form; in blue, the implied interval where $s_1 = 1$ along the annulus. We denote $\hat{p}_1 = \hat{p}(s_1 = 1)$. (b) Stretched annuluar augmentation; cf. Figure 1a.

choice is Rao-Blackwellized Gibbs sampling, which we will now explain.

Gibbs sampling in the annular augmentation is made possible by the fact that $\mathbf{s} = h(\mathbf{t}, \theta)$ is piecewise constant over the annulus (see braces in Figure 1a). In each of the intervals between threshold edges, constant $\mathbf{s} = h(\mathbf{t}, \theta)$ implies $\mathcal{L}(h(\mathbf{t}, \theta))$ is also constant. In Gibbs sampling the probability of sampling a point within an interval is proportional to the integral of the density over that interval. Here it is (proportional to) the product of the interval length and the value of $\mathcal{L}$ on the interval. Therefore we can Gibbs sample θ by first sampling an interval and then uniformly sampling a point within it. We further improve estimates from Gibbs sampling via Rao-Blackwellization (RB) (Douc et al., 2011). To calculate an expectation of some function $g(\mathbf{s})$, the RB estimator is:

$$\hat{\mathbb{E}}[g] = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{2d} r_k^n \cdot g(\mathbf{s}_k^n),$$

where r_k^n is the normalized density of the k^{th} interval between threshold edges in the n^{th} iteration and $\mathbf{s}_n^k$ is its corresponding value. Annular augmentation permits an efficient implementation of RB Gibbs sampling, which we show in pseudocode in Algorithm 1. As an implementation note, the rotational invariance of $\mathbf{t}$ and θ allows us to set $\theta = 0$ at the beginning of each iteration without loss of generality, thereby eliminating θ from Algorithm 1. Hereafter we will refer to Gibbs sampling on the annular augmentation as *Annular Augmentation Gibbs* sampling (AAG) and it's RB version as *Rao-Blackwellized Annular Augmentation Gibbs* sampling (AAG + RB).

Even though the AAG sampler may seem to involve more work than a traditional method such as MH, both have similar computational complexity per sample. To see this, note that each MH iteration requires evaluating f with a change in sign in one coordinate and a uniform random variable for acceptance-rejection. In AAG we require d uniform random variables to sample $\mathbf{t}$, must sort the threshold edges and perform $2d$ function evaluations of f with a change in sign in one coordinate. Although this is more expensive, using RB we get $2d$ samples, hence the cost per sample is similar. For uniform priors it is possible to avoid the cost of sorting the threshold edges if we constrain $\mathbf{t}$ to lie on a lattice, i.e. $t_i = k_i\pi/d$ where $k_i \in \{1, ..., 2d\}$.

We noted previously that the annular augmentation can be sampled with methods other than Gibbs. Indeed, in our experiments we will compare AAG to a similarly defined Annular Augmentation Slice sampler (AAS) (Neal, 2003). Suwa and Todo (2010) recently proposed a non-reversible MCMC sampler for discrete spaces with state-of-the-art mixing times; this Suwa-Todo (ST) algorithm can also be used in Operator 2. In Section 3.1, we extensively explore the benefits of using different Operator 2 approaches on Ising models.

2.4 Choice And Effect Of The Prior $\hat{p}$

Annular augmentation requires a distribution $\hat{p}$ which, for optimal performance, should approximately match the marginals of p. A natural choice is Loopy Belief Propagation (LBP), a deterministic approximation which incurs minimal computational overhead. The use of deterministic priors to speed up samplers has been used before in MCMC algorithms (De Freitas et al., 2001; Fagan et al., 2016) as well as in importance sampling (Liu et al., 2015). Often it makes the samplers more efficient without significant extra computational cost, as we will see in Section 3.

Annular augmentation explores the subset defined by the annular thresholds $\mathbf{t}$, and the behavior of the sampler will depend significantly on the choice of prior $\hat{p}$. Presuming $\hat{p}$ approximately matches the true marginals of p, it is instructive to consider two extreme cases that demonstrate the quality and flexibility

Algorithm 1: (AAG) Rao-Blackwellized Annular Augmentation Gibbs sampling (AAG + RB)

Input : Unnormalized density $\mathcal{L}$, prior $\hat{p}$, number of iterations N, function g
Output : Monte Carlo approximation $\hat{\mathbb{E}}[g]$
Initialize: $\mathbf{s} \in \{-1,+1\}^d$; $\hat{\mathbb{E}}[g] \leftarrow 0$
for $n = 1,..,N$ **do**
  /* Operator 1 */
  for $i = 1,..,d$ **do**
    $t_i \leftarrow \frac{\pi}{2}(s_i - 1) + \pi\hat{p}(s_i) \cdot unif(-1,+1)$
    // Threshold value
    $e_i \leftarrow (t_i - \pi\hat{p}(s_i = +1), t_i + \pi\hat{p}(s_i = +1))$
    // Threshold edges
  end
  /* Operator 2 */
  $q \leftarrow$ sort_edge_indices$(e_1, e_2, ..., e_d)$
  // Order of coordinate flips on annulus
  $\ell \leftarrow$ length$(q; e_1, e_2, ..., e_d)$
  // Length of ordered intervals
  Initialize $\mathbf{r}^n \in \mathbb{R}^{2d}$ // Interval densities
  for k=1,...,2d **do** // Move around annulus
    $s_{q(k)} \leftarrow -s_{q(k)}$
    $\mathbf{s}_k^n \leftarrow \mathbf{s}$
    $r_k^n \leftarrow \ell(k) \cdot \mathcal{L}(\mathbf{s}_k^n)$
  end
  $\mathbf{r}^n \leftarrow \mathbf{r}^n / \|\mathbf{r}^n\|_1$ // Normalize probabilities
  $i \sim Multinomial(\mathbf{r}^n)$, $\mathbf{s} \leftarrow \mathbf{s}_i^n$ // Gibbs sample
  $\hat{\mathbb{E}}[g] \leftarrow \hat{\mathbb{E}}[g] + \frac{1}{N}\sum_{k=1}^{2d} r_k^n \cdot g(\mathbf{s}_k^n)$
end

of the annular augmentation: strongly non-uniform $\hat{p}$, where $\hat{p}(s_i = 1) \approx 1$; and uniform priors, where $\hat{p}(s_i = 1) \approx 1/2$ for all $i = 1, ..., d$.

In the non-uniform case, each threshold is stretched to cover nearly all of the annulus, inducing a constant $\mathbf{s}$ around the annulus except for d tiny regions where only one coordinate is flipped (with high probability). Our sampler reduces to an MH type algorithm where only one coordinate is flipped at a time. This is appropriate and desired: nearly all of the density in p will be at $\mathbf{s} = (+1, ..., +1)$ with most of the remaining density residing in adjacent $\mathbf{s}'$ that have only one negative signed coordinate. AAG in such a case would quickly move to $\mathbf{s} = (+1, ..., +1)$ guided by the LBP prior, and then only consider flipping one coordinate at a time.

On the other hand, for uniform priors the annular augmentation will do a form of overrelaxed sampling. In overrelaxed sampling, points proposed lie on the opposite side of a mode or mean of a distribution (Neal, 1998) which can suppress random walks. In the annular augmentation with uniform prior, we can observe from Figure 1a that points on opposite sides of the annulus have opposite signs, i.e. $\mathbf{s} = h(\theta, \mathbf{t}) = -h(\theta + \pi, \mathbf{t})$. Annular augmentation considers all such points and has a similar flavor as an overrelaxed sampler. This is particularly useful for pairwise binary distributions without bias where $f(\mathbf{s}) = \prod_{i<j} \exp(\beta_{ij} s_i s_j)$. In this case $\mathbb{E}_p(\mathbf{s}) = 0$ since the density is symmetric with respect to signs. Here annular augmentation samplers simultaneously explore points of opposite signs, resulting in an extremely accurate estimate of node marginals.

3 EXPERIMENTAL RESULTS

We evaluate our annular augmentation framework on Ising models and a real world Boltzmann Machine (BM) application. Ising models have historically been an important benchmark for assessing the performance of binary samplers (Newman et al., 1999; Zhang et al., 2012; Pakman and Paninski, 2013). We use a set-up similar to (Zhang et al., 2012) to interpolate between unimodal and more challenging multi-modal target distributions which allows us to compare the performance of each sampler in different regimes and understand the benefits of RB and the LBP prior. BMs, also known as Markov random fields, are a fundamental paradigm in machine learning (Ackley et al., 1985; Barber, 2012), with it's deep counterpart having found numerous applications (Salakhutdinov and Hinton, 2009). Inference in fully connect BMs is extremely challenging. In our experiments we consider the heart disease dataset (Edwards and Toma, 1985) as its small size enables exact computations which can be used to measure the error of each sampler.

We compare against other general purpose state-of-the-art binary samplers: Exact-HMC (Pakman and Paninski, 2013) and Coordinate Metropolis Hastings (CMH), a MH sampler which proposes to flip only one coordinate, s_i, per iteration. We note that for certain classes of binary distributions, specialized samplers have been developed, e.g. the Wolff algorithm (Newman et al., 1999); as annular augmentation is a general binary sampling scheme, we do not compare to such specialized methods. To make the comparisons particularly challenging, we developed a novel method to incorporate the LBP prior in the CMH sampler and further improved its performance by using discrete event simulation. We refer to this as CMH + LBP (details in Appendix).

To distinguish the contributions of annular augmentation, RB, and the LBP pseudoprior, we show results separately for the AAG sampler, its RB version (AAG + RB), and with the LBP prior (AAG + RB + LBP). As mentioned in Section 2.3, we also provide results for the Annular Augmentation Slice (AAS) and Suwa-Todo (AAST) sampling counterparts. We run all samplers

for a fixed numbers of density function ($\mathcal{L}$) evaluations. This choice is sensible since it dominates the run time cost of each sampler, ensuring a fair comparison between different methods. All samplers were started from the same initial point, and for the LBP approximation we used Matlab code (Schmidt, 2007). We report results for estimation of i) node marginals, ii) pairwise marginals, and iii) the normalizing factor (partition function) where the exact values were obtained by running the junction tree algorithm (Schmidt, 2007).

3.1 2D Ising Model

We consider a 2D Ising model on a square lattice of size 9×9 with periodic boundary conditions, where $p(\mathbf{s}) \propto \exp\big(-\beta E[\mathbf{s}]\big)$. Here $E[\mathbf{s}] = -\sum_{\langle i,j \rangle} W_{ij} s_i s_j - \sum_i b_i s_i$ is the energy of the system and the sum $\sum_{\langle i,j \rangle}$ is over all adjacent nodes on the square lattice. The strength of interaction between node pairs (i, j) is denoted by W_{ij}, b_i is the bias applied to each node, and β is the inverse temperature of the system. As is often done, we fix $\beta = 1$, $W_{ij} = W$ for all (i, j) pairs, and apply a scaled bias $b_i \sim c \cdot Unif[-1, 1]$ to each node. Each MCMC method is run 20 times for 1000 equivalent density function ($\mathcal{L}$) evaluations.

In Figures 3, 4, and 5, we report RMSE of node marginal and pairwise marginal estimates, each averaged over 20 runs for different settings of W (referred to as "Strength") and c (referred to as "Bias scale"). A higher value in heatmaps (more yellow, less red) indicates a larger error. The values of strength and bias scale determine the difficulty of the problem, and demonstrate the performance of annular augmentation across a range of settings.

In Figure 3, we fix $c = 0.2$ and progressively increase W. These correspond to hard cases or "frustrated systems" with the target distribution being multimodal: increasing W increases the energy barrier between modes making the target more difficult to explore. All samplers do similarly well for easy, low-strength cases (left columns of Figure 3), but AAS, AAS + RB, AAG, AAG + RB and AAST, AAST + RB outperform on difficult problems (high W values, redder colors). Also interesting to note is that, in the high strength regime, the addition of LBP hurts performance (LBP itself, not annular augmentation, has broken down; see Appendix, Table 1 for numerical values).

In Figure 4, no bias is applied to any node and we progressively increase W. This corresponds to a target distribution with two modes: the ones and negative ones vectors. Samplers based on annular augmentation significantly outperform (lower error, more red) due to the overrelaxation property. For the zero bias cases, the LBP approximation is equal to the true node marginals: $\hat{p}(s_i = 1) = 0.5$ which is equivalent to the uniform prior case, hence we expect annular augmentation with and without LBP to perform similarly. A significant performance gain in estimating node marginals comes from RB (Appendix, Table 9 has numerical values). For settings in Figures 3 and 4, HMC, CMH and CMH + LBP samplers tend to get stuck in one mode, leading to poor performance.

Finally in Figure 5, we fix $W = 0.2$ and progressively increase c making the target distribution unimodal. In this simpler setting, HMC and CMH perform well and are on par with AAS + RB + LBP, AAG + RB + LBP and AAST + RB + LBP. We see this performance is due in part to the accuracy of LBP in this setting (Appendix, Table 2), shown also by the underperformance of annular augmentation without LBP in this case.

The annular augmented Gibbs, slice and ST methods have similar performance in all experiments. Gibbs always performs marginally better than slice and typically outperforms ST when RB. This suggests that Gibbs makes better use of RB, even though its mixing time may be slower (Suwa and Todo, 2010).

Regarding the size of these problems, note that annular augmentation has no issue scaling to much larger sizes; we stopped at 9×9 grids to be able to create the baseline with junction tree within a day or so of computation.

3.2 Heart Disease Risk Factors

The heart disease dataset (Edwards and Toma, 1985) lists six binary features (risk factors) relevant for coronary heart disease in 1841 men. These factors are modeled as a fully connected BM, which defines a probability distribution over a vector of binary variables $\mathbf{s} = [s_1, \ldots, s_n]$, $s_i \in \{0, 1\}$:

$$p(\mathbf{s}|\mathbf{W}, \mathbf{b}) = \frac{1}{Z(\mathbf{W}, \mathbf{b})} \exp\left\{ \sum_{i<j} W_{i,j} s_i s_j + \sum_i b_i s_i \right\}$$

A symmetric weight matrix $\mathbf{W}$, with zeros only along the diagonal, and a bias vector $\mathbf{b}$ parameterize the distribution with $W_{i,j}$ denoting the interaction strength between units i, j. Given a data set of binary vectors $\mathcal{D} = \{\mathbf{s}_j, \ \ j = 1, \ldots, N\}$, our goal is to learn the posterior over parameters $(\mathbf{W}, \mathbf{b})$. Given some prior $p(\mathbf{W}, \mathbf{b})$ we define the joint model as

$$p(\mathbf{W}, \mathbf{b}, \mathcal{D}) = p(\mathbf{W}, \mathbf{b}) \cdot p(\mathcal{D}|\mathbf{W}, \mathbf{b})$$

$$p(\mathcal{D}|\mathbf{W}, \mathbf{b}) = \frac{\exp\left\{ \sum_{n,i<j} W_{i,j} s_i^{(n)} s_j^{(n)} + \sum_{n,i} b_i s_i^{(n)} \right\}}{Z(\mathbf{W}, \mathbf{b})^N}$$

Bayesian learning here is hard: consider a simple MH sampler where starting from $(\mathbf{W}, \mathbf{b})$, a symmetric proposal distribution is used to propose $(\mathbf{W}', \mathbf{b}')$ which is

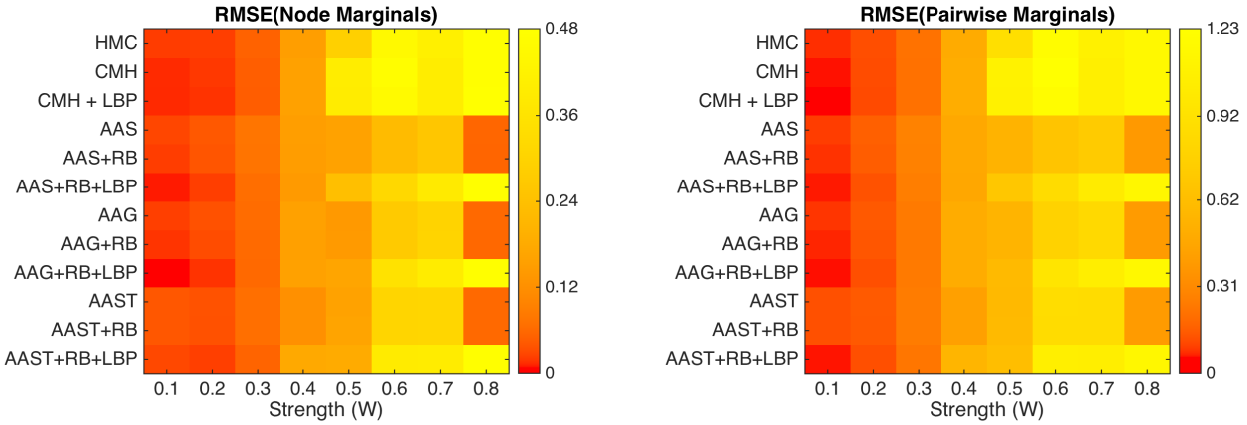


Figure 3: Bias scale $c = 0.2$ and the bias for each node is drawn as $b_i \sim c \cdot Unif[-1, 1]$. The target distribution here is multi-modal. As we increase strength W, AAS, AAS + RB and AAG, AAG + RB increase their outperformance over all other samplers, including those using LBP (the LBP approximate degrades for $W > 0.6$).

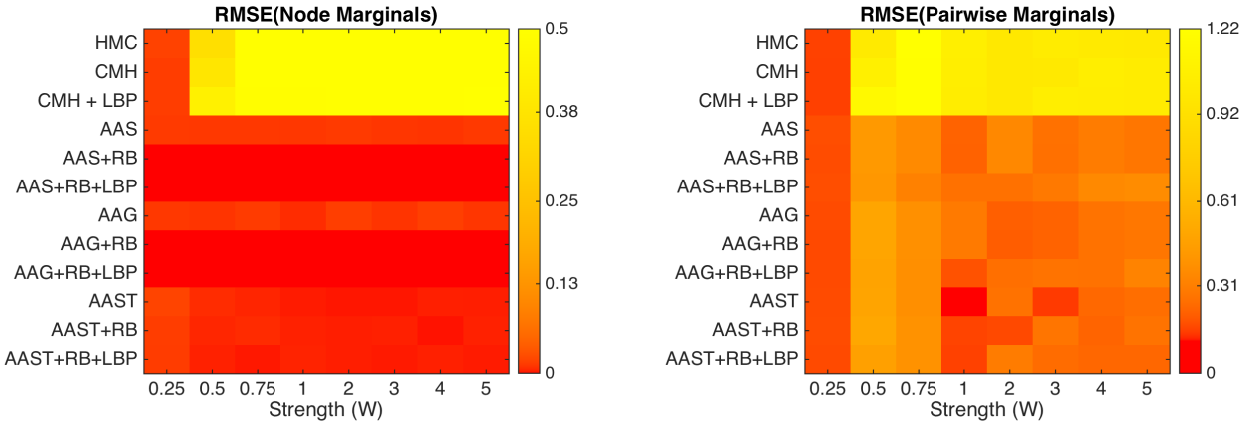


Figure 4: No bias is applied to any node making the target distribution bimodal, with the modes becoming more peaked as strength W is increased. All annular augmentation samplers very significantly outperform.

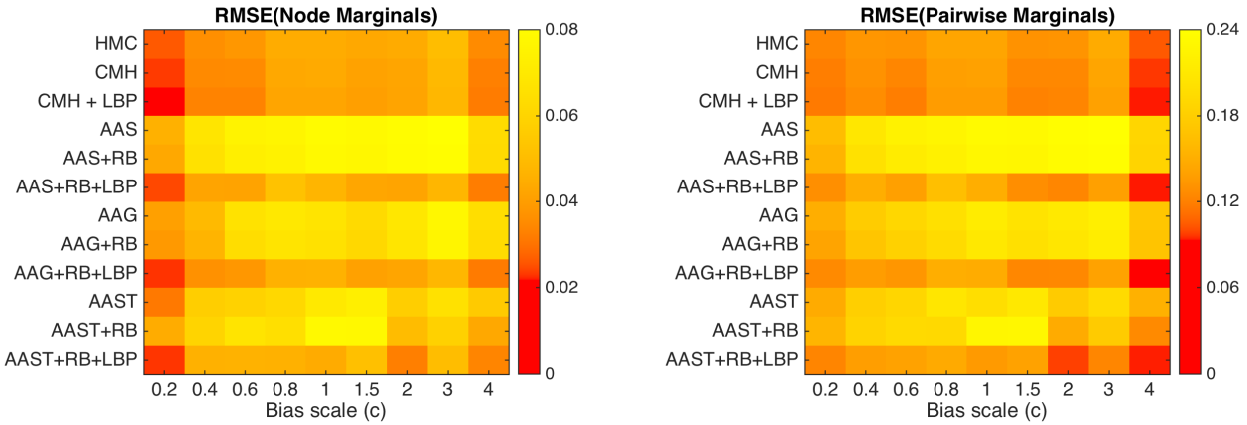


Figure 5: W is fixed to 0.2 and bias scale c is increased making the target distribution uni-modal. HMC and CMH find the mode quickly, as do annular augmentation samplers that leverage LBP. That group outperforms annular augmentation samplers with no access to LBP.

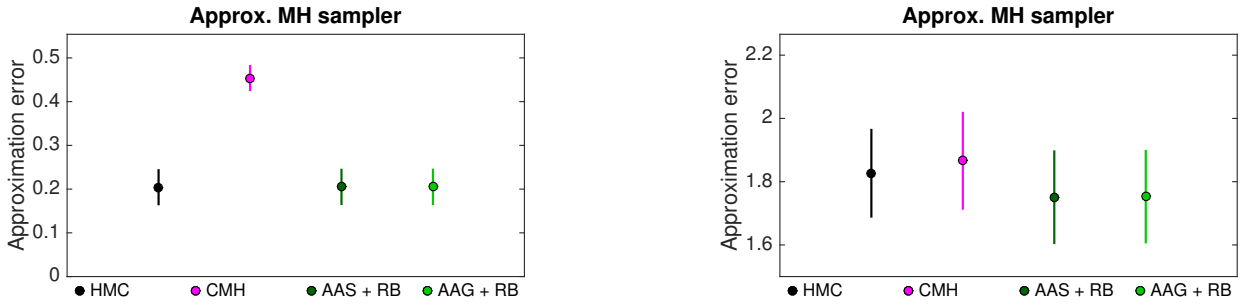


Figure 6: Absolute error in estimating the log partition function ratio for a) real data (left) and b) fake data simulation (right). For zero bias, LBP prior yields no additional benefit; we omit results for LBP driven samplers.

accepted with probability:

$$a = \min\left(1, \frac{p(\mathbf{W}', \mathbf{b}')}{p(\mathbf{W}, \mathbf{b})} \frac{p(\mathcal{D}|\mathbf{W}', \mathbf{b}')}{p(\mathcal{D}|\mathbf{W}, \mathbf{b})}\right), \tag{4}$$

which depends on the intractable partition function ratio $\left(Z(\mathbf{W}, \mathbf{b})/Z(\mathbf{W}', \mathbf{b}')\right)^N$. As in Murray and Ghahramani (2004), we use MCMC to approximate:

$$\begin{aligned} \frac{Z(\mathbf{W}, \mathbf{b})}{Z(\mathbf{W}', \mathbf{b}')} &\approx \frac{\tilde{Z}(\mathbf{W}, \mathbf{b})}{\tilde{Z}(\mathbf{W}', \mathbf{b}')} \\ &= \left\langle \exp\left\{ \sum_{i<j} (W'_{i,j} - W_{i,j}) s_i s_j + \sum_i (b'_i - b_i) s_i \right\} \right\rangle \end{aligned} \tag{5}$$

where $\langle \cdot \rangle$ denotes $E_{p(\mathbf{s}|\mathbf{W},\mathbf{b})}(\cdot)$. We will use different MCMC methods to draw samples from $p(\mathbf{s}|\mathbf{W}, \mathbf{b})$ to compute the expectation in Equation (5), which is then used to compute the approximate MH acceptance probability as in Equation (4). Following Murray and Ghahramani (2004), bias terms are taken to be zero.

We run 1000 such Markov chains, each with 100 samples (drawn using the approx. MH scheme). To approximate the partition function ratio in Equation (5), we use 1000 samples from 4 different samplers: i) HMC ii) CMH iii) AAS + RB and iv) AAG + RB. For each Markov chain, we compute the average absolute error in approximating the log partition function ratio, i.e.

$$\left\langle \left| \log\left(\frac{Z(\mathbf{W}, \mathbf{b})}{Z(\mathbf{W}', \mathbf{b}')}\right) - \log\left(\frac{\tilde{Z}(\mathbf{W}, \mathbf{b})}{\tilde{Z}(\mathbf{W}', \mathbf{b}')}\right) \right| \right\rangle_{p(\mathbf{W},\mathbf{b}|\mathcal{D})}$$

which is then averaged over 1000 such Markov chains. Following Murray and Ghahramani (2004), to see the effect of increasing dimensionality we also simulate data for 10 dimensional risk factors, using a random $\mathbf{W}$ matrix and repeat the experiment (details are given in the Appendix). Figure 6 plots the overall mean and standard deviation of the approximation error for both real and fake data. The exact partition function ratio, required for computing the error metric above, was evaluated by enumeration. For the real data example, both AAS + RB and AAG + RB outperform CMH and perform on par with HMC; while for the higher-dimensional fake data simulation, annular augmentation outperforms both CMH and HMC. Accordingly, across both real and simulated data, annular augmentation provides optimal performance. We note that small differences in Figure 6 can affect the convergence of the approximate MH sampler to the correct posterior distribution; as the approximation error for MH acceptance probabilities is proportional to the approximation error for partition function ratio (we plotted log of this quantity) taken to the N^{th} power.

4 CONCLUSION

We have presented a new framework for sampling from binary distributions. The annular augmentation sampler with subsequent Rao-Blackwellization has a number of desirable properties, including overrelaxation, no tuning parameters, and the ability to easily incorporate approximate posterior marginal information such as that from LBP. Taken together, these advantages lead to significant performance gains over CMH and Exact-HMC samplers, across a range of settings.

In this work we only considered uniform sampling of the thresholds: $t_i \sim unif(0, 2\pi)$; a future direction is to group thresholds together to flip clusters of correlated coordinates in the spirit of the Wolff algorithm. Furthermore, our inclusion of LBP in both annular augmentation and CMH suggests a similar extension for Exact-HMC (Pakman and Paninski, 2013).

Acknowledgements

We would like to thank the anonymous reviewers, especially for suggesting the Suwa-Todo method, as well as Ari Pakman and Liam Paninski for helpful discussions.

FF is supported by a grant from Bloomberg. JPC is supported by Simons Foundation (SCGB#325171 and SCGB#325233), the Sloan Foundation, the McKnight Foundation, and the Grossman Center.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Chakraborty, S., Fremont, D. J., Meel, K. S., Seshia, S. A., and Vardi, M. Y. (2014). Distribution-aware sampling and weighted model counting for SAT. *arXiv preprint arXiv:1404.2984*.

Chandrasekaran, V., Srebro, N., and Harsha, P. (2008). Complexity of inference in graphical models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 70–78. AUAI Press.

Chavira, M. and Darwiche, A. (2008). On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6):772–799.

De Freitas, N., Højen-Sørensen, P., Jordan, M. I., and Russell, S. (2001). Variational MCMC. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 120–127. Morgan Kaufmann Publishers Inc.

Douc, R., Robert, C. P., et al. (2011). A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *The Annals of Statistics*, 39(1):261–277.

Edwards, D. and Toma, H. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351.

Fagan, F., Bhandari, J., and Cunningham, J. P. (2016). Elliptical slice sampling with expectation propagation. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo.

Liu, Q., Fisher III, J. W., and Ihler, A. T. (2015). Probabilistic variational bounds for graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc.

Murray, I. and Ghahramani, Z. (2004). Bayesian learning in undirected graphical models: approximate mcmc algorithms. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 392–399. AUAI Press.

Murray, I., Prescott, R., David, A., and Mackay, J. C. (2010). Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Neal, R. (2003). Slice sampling. *Annals of Statistics*, 31:705–741.

Neal, R. M. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. In *Learning in graphical models*, pages 205–228. Springer.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2.

Newman, M. E., Barkema, G. T., and Newman, M. (1999). *Monte Carlo methods in statistical physics*, volume 13. Clarendon Press Oxford.

Nowozin, S. and Lampert, C. H. (2011). Structured prediction and learning in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4):3–4.

Pakman, A. and Paninski, L. (2013). Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. In *Advances in Neural Information Processing Systems*, pages 2490–2498.

Salakhutdinov, R. and Hinton, G. E. (2009). Deep boltzmann machines. In *AISTATS*, volume 1, page 3.

Schmidt, M. (2007). UGM: A Matlab toolbox for probabilistic undirected graphical models.

Suwa, H. and Todo, S. (2010). Markov chain Monte Carlo method without detailed balance. *Physical review letters*, 105(12):120603.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Wang, F. and Landau, D. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050.

Zhang, Y., Ghahramani, Z., Storkey, A. J., and Sutton, C. A. (2012). Continuous relaxations for discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 3194–3202.