Exact and Approximate Sampling by Systematic Stochastic Search

Vikash Mansinghka
MIT BCS & CSAIL

Daniel Roy
MIT CSAIL

Eric Jonas
MIT BCS

Joshua Tenenbaum
MIT BCS & CSAIL

Abstract

We introduce *adaptive sequential rejection sampling*, an algorithm for generating exact samples from high-dimensional, discrete distributions, building on ideas from classical AI search. Just as systematic search algorithms like A* recursively build complete solutions from partial solutions, sequential rejection sampling recursively builds exact samples over high-dimensional spaces from exact samples over lower-dimensional subspaces. Our algorithm recovers widely-used particle filters as an approximate variant without adaptation, and a randomized version of depth first search with backtracking when applied to deterministic problems. In this paper, we present the mathematical and algorithmic underpinnings of our approach and measure its behavior on undirected and directed graphical models, obtaining exact and approximate samples in a range of situations.

1 Introduction

Efficient inference in high-dimensional, discrete probability models is a central problem in computational statistics and probabilistic AI. In this paper, we introduce a recursive algorithm for exact sampling aimed at solving this problem in the presence of multimodality. We do this by generalizing ideas from classic AI search to the stochastic setting. Just as systematic search algorithms like A* recursively build complete solutions from partial solutions, sequential rejection sampling recursively builds exact samples over high-dimensional spaces from exact samples over lower-dimensional subspaces. Our method exploits and generalizes ideas from classical AI search for managing deterministic dependencies, including depth first traversal with early constraint checking and backtracking, to tractably generate exact and approximate samples.

Many popular approaches to inference, such as mean-field variational methods (Jordan et al., 1999), convex relaxations (Wainwright et al., 2002; Sontag and Jaakkola, 2008), and generalized belief propagation (Yedidia et al., 2001), focus on approximating MAP assignments or (low-dimensional, e.g. 1 or 2 variable) marginal distributions. While effective in many settings, low-dimensional marginals (and MAP assignments) often do not capture the essential features of a distribution, especially in the presence of multimodality. Homogeneous Ising models, where the probability of a state $x = (x_1, \ldots, x_n)$ is[1]

$$P(\mathbf{x}) \propto \exp\Big\{-J \sum_{(i,j)\in E} x_i x_j\Big\}, \quad \mathbf{x} \in \{-1, 1\}^n, \qquad (1)$$

provide one source of extreme examples. As the coupling parameter J increases, the joint distribution on spins approaches a 2-component mixture on the "all up" and "all down" states, which has only 1 bit of entropy. A MAP approximation misses the fundamental bimodality of the distribution, while the minimum-KL product-of-marginals approximation confuses this distribution with the uniform distribution on spins.

When a distribution contains many widely separated modes, and is therefore difficult to parametrically approximate, simulation-based inference seems ideal. Exact and approximate sampling are also of intrinsic interest in computational physics (Propp and Wilson, 1996; Edwards and Sokal, 1988). Unfortunately, popular methods like Gibbs sampling often run into

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

[1] We use $f(x) \propto g(x, y)$ to mean that $f(x) = c(y)\, g(x, y)$ for some (in the case of distributions, normalizing) constant of proportionality $c(y)$.

severe convergence problems in precisely these settings. This difficulty motivates a number of specialized samplers that exploit sophisticated data augmentation techniques (Swendsen and Wang, 1987), as well as a variety of model-specific proposal designs.

Our algorithm mitigates the problems of multimodality by generalizing ideas for managing deterministic dependencies from the constraint satisfaction literature. In particular, it operates by a stochastic generalization of *systematic* search. In deterministic systematic search, solutions to a problem are built up piece-by-piece[2]. The first complete candidate solution is either exact (as in backtracking search or A*) or approximate (as in beamed searches), and strategies for search tree expansion are often used to manage deterministic dependencies among chains of choices. Our algorithm automatically recovers a randomized variant of one such method, depth first search with backtracking and early constraint checking, when applied to constraint satisfaction problems, generalizing these ideas to the setting of sampling. Furthermore, if the rejection step in our algorithm is replaced with importance sampling with resampling (and particular restricted choices of variable orderings are used) we recover widely-used particle filtering algorithms for approximate sampling.

In this paper, we present the mathematical and algorithmic underpinnings of our approach and measure its behavior on Ising models, causal networks and a stereo vision Markov random field, obtaining exact and approximate samples in a range of situations.

2 The Adaptive Sequential Rejection Algorithm

Consider the problem of generating samples from an high-dimensional discrete distribution with density $P(\mathbf{x})$, $\mathbf{x} \in \mathbf{X}$. In many cases, we can only efficiently compute the density to within a multiplicative constant; that is, we can compute a function $\bar{P}(\mathbf{x})$ such that $P(\mathbf{x}) = \bar{P}(\mathbf{x})/Z$. This setting arises in factor graph inference, where the normalization constant $Z = \sum_{\mathbf{x}\in\mathbf{X}} \bar{P}(\mathbf{x})$ is due to either the partition function of the underlying undirected model or the marginal likelihood of the evidence in the underlying directed model. Let $\bar{P}(\mathbf{x}) = \bar{P}(\mathbf{y}, z) = \psi_1(\mathbf{y})\,\psi_2(\mathbf{y}, z)$, where $\mathbf{X} = \mathbf{Y} \times Z$ is a decomposition of the state that results in a factored representation for $\bar{P}$. For example, we might take Z to be one variable, and $\mathbf{Y}$ to be all the rest. Our algorithm generates exact samples from $P(\mathbf{y}, z)$ by recursively generating an exact sample $\hat{\mathbf{y}}$ from $\bar{P}'(\mathbf{y}) = \psi_1(\mathbf{y})$ (which we assume has an analogous decomposition, i.e. $\mathbf{Y}$ and ψ_1 factor), and then extending $\hat{\mathbf{y}}$ to an exact sample $(\hat{\mathbf{y}}, \hat{z})$ from $P(\mathbf{y}, z)$ by rejection.

In order to apply our algorithm to an arbitrary factor graph, we will need a way of recursively decompose a model into a nested sequence of distributions. We will return to this issue later in the paper, using ideas from deterministic search, where variables will be brought in one at a time. First, we present the basic recursion in our algorithm, assuming a decomposition has been chosen.

Assume (by induction) that we have an exact sample $\hat{\mathbf{y}}$ from $P'(\mathbf{y})$. Let $\bar{P}(\mathbf{y}) = \sum_{z'} \bar{P}(\mathbf{y}, z')$ be the (unnormalized) marginal distribution of $\mathbf{y}$ under P. We define[3] a (Gibbs) transition kernel from $\mathbf{Y}$ into Z by

$$\begin{aligned} Q_P(z \mid \mathbf{y}) &\triangleq \frac{\bar{P}(\mathbf{y}, z)}{\bar{P}(\mathbf{y})} && (2) \\ &= \frac{\psi_1(\mathbf{y})\,\psi_2(\mathbf{y}, z)}{\sum_{z'} \psi_1(\mathbf{y})\,\psi_2(\mathbf{y}, z')} && (3) \\ &= \frac{\psi_2(\mathbf{y}, z)}{\sum_{z'} \psi_2(\mathbf{y}, z')}, && (4) \end{aligned}$$

and sample $\hat{z}$ from $Q_P(\,\cdot \mid \hat{\mathbf{y}})$. We then treat $(\hat{\mathbf{y}}, \hat{z})$, whose density is $P'(\mathbf{y})\,Q_P(z \mid \mathbf{y})$, as a proposal to a rejection sampler for $P(\mathbf{y}, z)$. Let $\hat{\mathbf{x}} = (\hat{\mathbf{y}}, \hat{z})$ and define the weight of the sample $\hat{\mathbf{x}}$ as

$$\begin{aligned} W_{P' \to P}(\hat{\mathbf{x}}) &\triangleq \frac{\bar{P}(\hat{\mathbf{y}}, \hat{z})}{\bar{P}'(\hat{\mathbf{y}})\,Q_P(\hat{z} \mid \hat{\mathbf{y}})} && (5) \\ &= \frac{\bar{P}(\hat{\mathbf{y}})}{\bar{P}'(\hat{\mathbf{y}})} && (6) \\ &= \frac{\sum_{z'} \psi_1(\hat{\mathbf{y}})\,\psi_2(\hat{\mathbf{y}}, z')}{\psi_1(\hat{\mathbf{y}})} && (7) \\ &= \sum_{z'} \psi_2(\hat{\mathbf{y}}, z'). && (8) \end{aligned}$$

Note that the weight does not depend on $\hat{z}$ and so we consider the weight $W_{P' \to P}(\mathbf{y})$ a function of $\mathbf{y}$.

We then accept $\hat{\mathbf{x}}$ as an exact sample from P with probability

$$\frac{W_{P' \to P}(\hat{\mathbf{y}})}{C_{P' \to P}}, \qquad (9)$$

[2] Contrast with *local* search, e.g. fixed-point iteration, where a complete but possibly poor quality approximate solution is repeatedly iterated upon until it stabilizes. Markov chain Monte Carlo methods generalize this fixed-point iteration idea to the space of distributions, asymptotically converging to the correct distribution and recovering deterministic fixed-point algorithms for particular proposal and target choices. Techniques like coupling from the past (Propp and Wilson, 1996; Huber, 2002; Childs et al., 2000) provide the distributional analogues of termination analysis, sometimes allowing automatic determination of when exact samples have been obtained.

[3] We will use $\triangleq$ to denote definitions.

where $C_{P'\to P}$ is any constant such that $C_{P'\to P} \geq W_{P'\to P}(\mathbf{y})$ for all $\mathbf{y} \in \mathbf{Y}$. In general, loose upper bounds on $W_{P'\to P}(\mathbf{y})$ are easy to find but introduce unnecessary rejections. On the other hand, overconfident values of C are also easy to find, but will result in approximate samples (or, more precisely, exact samples from the distribution proportional to $\min\{\bar{P}(\mathbf{y}, z), C_{P'\to P}\, \bar{P}'(\mathbf{y})\, Q_P(z \mid \mathbf{y})\}$.) Both variants may have practical value. Here, we focus on the setting where we actually use the optimal rejection constant:

$$C^*_{P'\to P} \triangleq \max_{\mathbf{y}} W_{P'\to P}(\mathbf{y}) = \max_{\mathbf{y}} \sum_{z'} \psi_2(\mathbf{y}, z'). \quad (10)$$

If $\mathbf{y} = (y_1, \ldots, y_n)$ is high-dimensional, then the worst case complexity of calculating $C^*_{P'\to P}$ is exponential in n. However, when the sequence of distributions we are using has sparse dependencies (i.e., when $\psi_2(y, z)$ is a function of only $O(\log n)$ dimensions y_i), then we can calculate $C^*_{P'\to P}$ in polynomial time. For example, in 2-d grid Ising models, ψ_2 depends on at most three neighbors and therefore C^* can be calculated in constant time.

This inductive argument describes the non-adaptive sequential rejection sampler. We apply it to sampling from the joint distribution of factor graph models by automatically constructing a nested sequence of distributions from orderings on the variables, using machinery we introduce later. Sequential importance sampling with resampling (SIR) can be viewed as an approximate variant, where the rejection step - producing one exact sample, or failing - is replaced with an importance sampling and resampling step propagating k particles approximately drawn from the target. The weight per particle is the same as in sequential rejection.

The choice of the Gibbs transition kernel is important. Incorporating the $\psi_2(\mathbf{y}, z)$ factor into the proposal prevents the algorithm from proposing samples $\hat{z}$ that are already known to be incompatible with the setting $\hat{\mathbf{y}}$. Thus we recover early constraint checking, and generalize it to favor paths that seem probable given the current partial assignment.

2.1 Adaptation Stochastically Generalizes Backtracking

Systematic searches typically avoid reconsidering partial solutions that have been discovered inconsistent; this behavior is known as backtracking, and requires dynamically recording the information about inconsistent states obtained over the course of search. We accomplish this in the broader setting of sampling by introducing an adaptation rule into our sampler, which recovers this deterministic avoidance in the limit of deterministic inconsistency.

Following Eq. 9, the non-adaptive sampler with the optimal rejection constants C^* accepts samples with probability

$$\alpha_{P'\to P} = \frac{\mathbb{E}_{P'}(W_{P'\to P}(\hat{\mathbf{y}}))}{C^*_{P'\to P}}. \quad (11)$$

From Eq. 6, we have that

$$W_{P'\to P}(\mathbf{y}) \propto \frac{P(\mathbf{y})}{P'(\mathbf{y})}, \quad (12)$$

and therefore, using the definition of $C^*_{P'\to P}$ and canceling the constant of proportionality shared between W and C^*, we have

$$\alpha_{P'\to P} = \frac{\sum_{\mathbf{y}} P'(\mathbf{y}) \frac{P(\mathbf{y})}{P'(\mathbf{y})}}{\max_{\mathbf{y}} \frac{P(\mathbf{y})}{P'(\mathbf{y})}} \quad (13)$$

$$= \min_{\mathbf{y}} \frac{P'(\mathbf{y})}{P(\mathbf{y})}. \quad (14)$$

Note that the acceptance probability $\alpha_{P'\to P}$ depends only on the choice of P' and P and is precisely the smallest ratio in probability assigned to some $y \in \mathbf{Y}$.[4] An interesting special case is when the simpler distribution $P'(\mathbf{y})$ matches the marginal $P(\mathbf{y})$. In this case, $W_{p'\to p} = 1$ and we always accept.[5] Assuming each attempt to generate samples from P' by rejection succeeds with probability $\alpha_{P'}$, the entire rejection sampler will succeed with probability $\alpha_{P'}\alpha_{P'\to P}$. If this probability is $O(2^{-w})$, where w is the tree width of the factor graph, then, in expectation, we will be no better off than using variable clustering and dynamic programming to calculate marginals and sample exactly.

Our goal then is to drive $\alpha_{P'\to P} \to 1$ (and inductively, $\alpha_{P'} \to 1$). Consider the extreme case where a sampled value $\hat{\mathbf{y}}$ is revealed to be inconsistent. That is, $\psi_2(\hat{\mathbf{y}}, z) = 0$ for all z and therefore $W_{P'\to P}(\mathbf{y}) = 0$. We should then adjust P' (and its predecessors, recursively) so as to never propose the value $\mathbf{y} = \hat{\mathbf{y}}$ again. Certainly if P' is the marginal distribution of P (recursively along the chain of rejections), this will take place.

Consider the (unnormalized) proposal density

$$\bar{P}'_S(\mathbf{y}) = \bar{P}'(\mathbf{y}) \prod_{\mathbf{y}' \in S} \left(\frac{W_{P'\to P}(\mathbf{y}')}{C^*_{P'\to P}} \right)^{\delta_{\mathbf{y}\mathbf{y}'}} \quad (15)$$

[4] In particular, the acceptance is positive only if $P(\mathbf{y}) > 0 \implies P'(\mathbf{y}) > 0$ (i.e., P' is absolutely continuous with respect to P).

[5] While it may be tempting to think the problem is solved by choosing $P = P'$, if each stage of the algorithm performed this marginalization, the overall complexity would be exponential. The key to adaptation will be selective feedback.

where $S \subset \mathbf{Y}$ and $\delta_{\mathbf{y}\mathbf{y}'}$ is the Kronecker delta satisfying $\delta_{\mathbf{y}\mathbf{y}'} = 1$ if $\mathbf{y} = \mathbf{y}'$ and 0 otherwise. Then

$$W_{P'_S \to P}(\mathbf{x}) \triangleq \frac{\bar{P}(\mathbf{y}, z)}{\bar{P}'_S(\mathbf{y})\, Q_P(z \mid \mathbf{y})} \tag{16}$$

$$= \frac{W_{P' \to P}(\mathbf{y})}{\prod_{\mathbf{y}' \in S} \left(\frac{W_{P' \to P}(\mathbf{y}')}{C^*_{P' \to P}} \right)^{\delta_{\mathbf{y}\mathbf{y}'}}} \tag{17}$$

$$= \begin{cases} C^*_{P' \to P} & \mathbf{y} \in S \\ W_{P' \to P}(\mathbf{y}) & \mathbf{y} \notin S, \end{cases} \tag{18}$$

where step (17) follows from Eq. (6). Therefore $C^*_{P'_S \to P} = C^*_{P' \to P}$. In particular, if $S = \mathbf{Y}$, then $W_{P'_S \to P}(\mathbf{y}) = C^*_{P'_S \to P} = C^*_{P' \to P}$ and every sample is accepted. In fact,

$$\bar{P}'_{S=\mathbf{Y}}(\mathbf{y}) = \bar{P}'(\mathbf{y}) \prod_{\mathbf{y}' \in \mathbf{Y}} \left(\frac{W_{P' \to P}(\mathbf{y}')}{C^*_{P' \to P}} \right)^{\delta_{\mathbf{y}\mathbf{y}'}} \tag{19}$$

$$\propto \bar{P}'(\mathbf{y})\, W_{P' \to P}(\mathbf{y}) \tag{20}$$

$$= \bar{P}'(\mathbf{y})\, \frac{\bar{P}(\mathbf{y})}{\bar{P}'(\mathbf{y})} \tag{21}$$

$$= \bar{P}(\mathbf{y}) \tag{22}$$

and therefore an exact sample from $P'_{\mathbf{Y}}$ is a sample from the marginal distribution of P. The Gibbs kernel exactly extends this to a sample from the joint.

Adaptation then involves the following modification to our algorithm: after proposing a sample $(\hat{\mathbf{y}}, \hat{z})$, we augment S with $\hat{\mathbf{y}}$. As $S \to \mathbf{Y}$, $\bar{P}'_S(\mathbf{y}) \to \frac{1}{C^*_{P' \to P}} \bar{P}(\mathbf{y})$ pointwise.

This change can be implemented efficiently by storing a hashmap of visited states for every distribution in the sequence and modifying density evaluation (and, therefore, the Gibbs kernels) to reflect hashmap contents. Each stage of the sampler pushes states to the previous stage's hashmap as adaptation proceeds, moving each proposal distribution towards the ideal marginal. Because such adaptation leaves C^* unchanged (see Appendix), adaptation increases the algorithmic complexity by only a linear factor in the number of sampling attempts, with overall complexity per attempt still linear in the number of variables. Taken together, the hashmaps play the role of the stack in a traditional backtracking search, recording visited states and forbidding known bad states from being proposed.

2.2 Sequences of Distributions for Graphical Models

To apply this idea to graphical models, we need a way to generically turn a graphical model into a sequence of distributions amenable to adaptive sequential rejection. We accomplish this - and introduce further ideas from systematic search - by introducing the idea of a *sequence of restrictions* of a given factor graph, based on a *variable ordering* (i.e. permutation of the variables in the model). Each sequence of restrictions can be deterministically mapped to a nested sequence of factor graphs which, for many generic orderings, capture a good sequence of distributions for sequential rejection under certain analytically computable conditions.

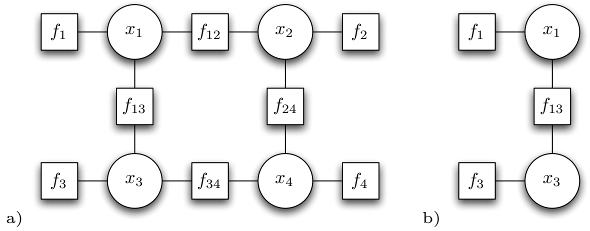

Figure 1: A four node Ising model, and its restriction to the variables x_1 and x_3.

We denote by X_i a random variable taking values $x_i \in \mathbf{X}_i$. If $V = (X_1, \ldots, X_k)$ is a vector random variables, then we will denote by $\mathbf{X}_V$ the cartesian product space $\mathbf{X}_1 \times \cdots \times \mathbf{X}_k$ in which the elements of V take values $v = (x_1, \ldots, x_k)$.

Definition 2.1 *A factor graph $G = (X, \Psi, V)$ is an undirected X, Ψ-bipartite graph where $X = (X_1, \ldots, X_n)$ is a set of random variable and $\Psi = \{\psi_1, \ldots, \psi_m\}$ is a set of factors. The factor ψ_i represents a function $\mathbf{X}_{V_i} \mapsto [0, \infty]$ over the variables $V_i \subset X$ adjacent to ψ_i in the graph. The graph represents a factorization*

$$P(v) = P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i \psi_i(v_i) \tag{23}$$

of the probability density function P, where Z is the normalization constant.

Definition 2.2 *The* restriction G_S of the factor graph $G = (X, \Psi, V)$ to a subset $S \subset X$ *is the subgraph (S, Ψ_S, V_S) of G consisting of the variables S, the collection of factors $\Psi_S = \{\psi_i \in \Psi \mid V_i \subset S\}$ that depend only on the variables S, and the edges $V_S = \{V_i \mid \psi_i \in \Psi_S\}$ connecting the variables S and factors Ψ_S. We denote by Z_S the normalization constant for the restriction.*

See Figure 1 for an example restriction of an Ising model. Consider a factor graph $G = (X, \Psi, V)$ and let $X_{1:k} = \{x_1, \ldots, x_k\} \subset X$, $(k = 1, \ldots, n)$ be the first k variables in the model under some order. The sequence of distributions we consider are the distributions given by the restrictions $G_{X_{1:k}}$, $k = 1, \ldots, n$.

We recover likelihood weighting (generalizing it to include resampling) on Bayesian networks when we use the importance variant of our algorithm and a topological ordering on the variables. Similarly, we recover particle filtering when we apply our method to time series models, with resampling instead of rejection and an ordering increasing in time.

In this paper, we focus on generically applicable strategies for choosing an ordering. All our exact sampling results use a straightforward ordering which first includes any deterministically constrained variables, then grows the sequence along connected edges in the factor graph (with arbitrary tie breaking). This way, as in constraint propagation, we ensure we satisfy known constraints before attempting to address our uncertainty about remaining variables. If we do not do this, and instead sample topologically, we find that unlikely evidence will lead to many rejections (and approximate rejection, i.e. likelihood weighting, will exhibit high variance). In general, we expect finding an optimal ordering to be difficult, although heuristic ordering information (possibly involving considerable computation) could be exploited for more efficient samplers. An adaptive inference planner, which dynamically improves its variable ordering based on the results of previous runs, remains an intriguing possibility.

3 Experiments

First, we measure the behavior on ferromagnetic Ising models for a range of coupling strengths, including the critical temperature and highly-coupled regimes where Gibbs samplers (and inference methods like mean-field variational and loopy belief propagation) have well-known difficulties with convergence; see Figure 3 shows some of our results.

We have also used our algorithm to obtain exact samples from 100x100-dimensional antiferromagnetic (repulsive) grid Ising models at high coupling, with no rejection, as is expected by analytic computation of the αs, describing probability of acceptance. At this scale, exact methods such as junction tree are intractable due to treewidth, but the target distribution is very low entropy and generic variable orderings that respect connectedness lead to smooth sequences and therefore effective samplers. We have also generated from exact samples from 20x20-dimensional ferromagnetic grid Isings at more intermediate coupling levels, where adaptation was critical for effective performance.

We also measured our algorithm's behavior on randomly generated (and in general, frustrated) Ising models with coupling parameters sampled from $U[-2, 2]$. We report results for a typical run of the adaptive and non-adaptive variants of sequential rejection sampling on a typical problem size; see Figure 4 for details. We also note that we have successfully obtained exact samples from 8x8-dimensional Isings with randomly generated parameters, using adaptation. On the models we tested, we obtained our first sample in roughly 5000 attempts, reducing to roughly one sample per 1000 attempts after a total of 100,000 had been made.

Given the symmetries in the pairwise potentials in (even) a frustrated Ising model without external field - the score is invariant to a full flip of all states - our algorithm will always accept with probability 1 on tree-structured (sub)problems. This is because the combination of Gibbs proposals (i.e. the arc-consistency insight) with the generic sequence choice (connected ordering) can always satisfy the constraints induced by the agreement or disagreement on spins in these

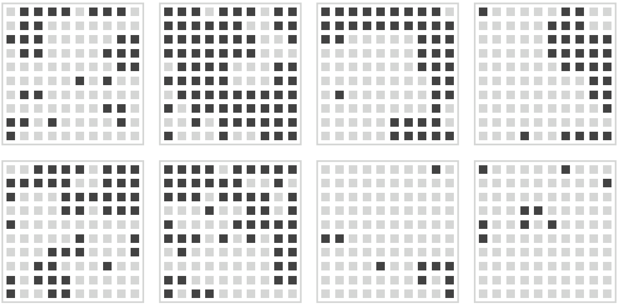

Figure 2: (left 4) Exact samples from a 10x10-dimensional grid ferromagnetic Ising just below the critical temperature. (right 4) Exact samples from a 10x10-dimensional grid ferromagnetic Ising just above the critical temperature.

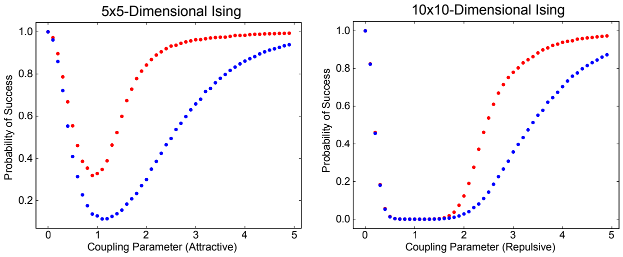

Figure 3: (left) Ferromagnetic (right) Antiferromagnetic. (both) Frequency of acceptance in nonadaptive (blue, lower) and adaptive (red, higher) sequential rejection as a function of coupling strength J. Naive rejection approaches suffer from exponential decay in acceptance probability with dimension across all coupling strengths, while generic MCMC approaches like Gibbs sampling fail to converge when the coupling reaches or exceeds the critical value. Note that adaptive rejection improves the bound on the region of criticality.

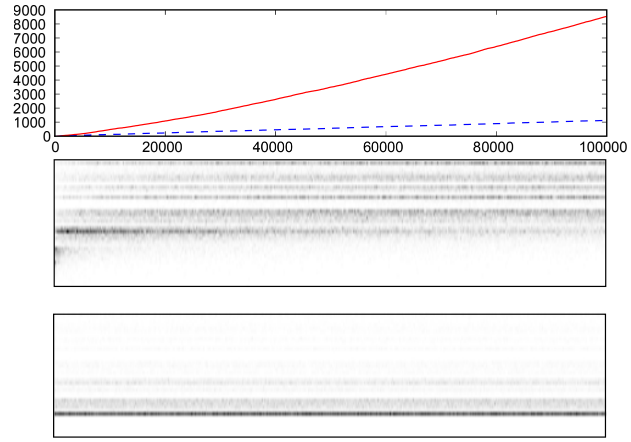

Figure 4: Comparison of adaptive (top-red, center) and nonadaptive (top-blue/dashed, bottom) rejection sampling on a frustrated 6x6-dimensional Ising model with uniform $[-2, 2]$ distributed coupling parameters. (top) Cumulative complete samples over 100,000 iterations. (lower plots) A black dot at row i of column j indicates that on the jth iteration, the algorithm succeed in sampling values for the first i variables. Only a mark in the top row indicates a successful complete sample. While the nonadaptive rejection sampler (bottom) often fails after a few steps, the adaptive sampler (center), quickly adapts past this point and starts rapidly generating samples.

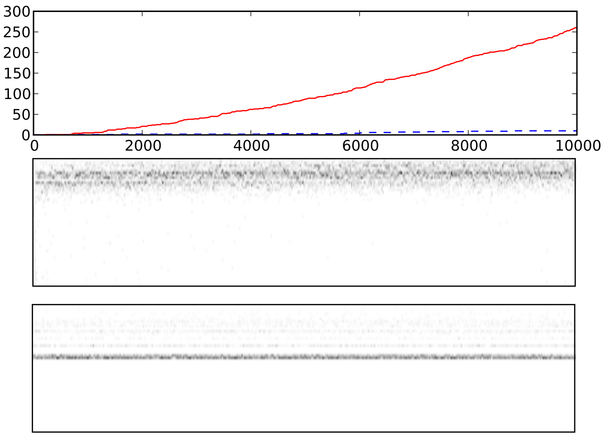

Figure 5: Comparison of adaptive (top-red, center) and nonadaptive (top-blue/dashed, bottom) rejection sampling for posterior inference on a randomly generated medical diagnosis network with 20 diseases and 30 symptoms. The parameters are described in the main text. (top) Cumulative complete samples over 100,000 iterations. (lower plots) show the trajectories of a typical adaptive and nonadaptive run in the same format as Figure 4. Here, adaptation is critical, as otherwise the monolithic noisy-OR factors result in very low acceptance probabilities in the presence of explaining away.

settings. Accordingly, our algorithm is more efficient than other exact methods for trees (such as forward filtering with backward sampling) in these cases. If, on the other hand, the target distribution does not contain this symmetry (so some of the initial choices matter), there will be some probability of rejection, unlike with forward filtering and backward sampling. This helps to explain the bands of rejection sometimes seen in the nonadaptive algorithm and the opportunity for adaptation on Ising models, as it is impossible for the algorithm to reject until a variable is added when its already added neighbors disagree.

We also applied our method to the problem of diagnostic reasoning in bipartite noisy-OR networks. These problems motivated several variational inference algorithms and in which topological simulation and belief propagation are known to be inaccurate (Saul et al., 1996). Furthermore, as in the ferromagnetic Ising setting, it seems important to capture the multimodality of the posterior. A doctor who always reported the most probable disease or who always asserted you were slightly more likely to be sick having visited him would not have a long or successful professional life. The difficulty of these problems is due to the rarity of diseases and symptoms and the phenomenon of "explaining away", yielding a highly multimodal posterior placing mass on states with very low prior probability. We explored several such networks, generating sets of symptoms from the network and measuring both the difficulty of obtaining exact samples from the full posterior distribution on diseases and the diagnostic accuracy. Figure 5 shows exact sampling results, with and without adaptation, for a typical run on a typical network, generated in this regime. This network had 20 diseases and 30 symptoms. Each possible edge was present with probability 0.1, with a disease base rate of 0.2, a symptom base rate of 0.3, and transmission probabilities of 0.4.

The noisy-OR CPTs result in large factors (with all diseases connected through any symptoms they share). Accordingly, the sequential rejection method gets no partial credit by default for correctly diagnosing a symptom until all values for all possible diseases have been guessed. This results in a large number of rejections. Adaptation, however, causes the algorithm to learn how to make informed partial diagnoses better and better over exposure to a given set of symptoms.

Finally, we applied our method to a larger-scale application: approximate joint sampling from a Markov Random Field model for stereo vision, using parameters from (Tappen and Freeman, 2003). This MRF had 61,344 nodes, each with 30 states; Figure 6 shows our results. We suspect exact sampling is intractable at this scale, so we used the importance relaxation of our algorithm with 1 and 5 particles. Because of the strong - but not deterministic - influence of the

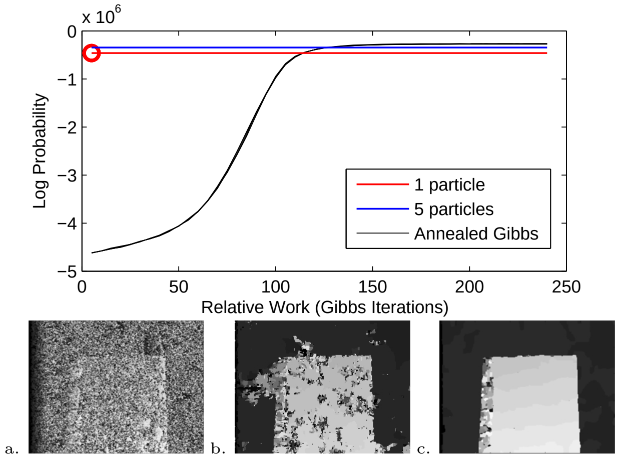

Figure 6: Comparison of an aggressively annealed Gibbs sampler (linear temperature schedule from 20 to 1 over 200 steps) to the non-adaptive, importance relaxation of our algorithm. The red circle denotes the mean of three 1-particle runs. The horizontal bars highlight the quality of our result. (a) Gibbs stereo image after sampling work comparable to an entire 1-particle pass of (b) our algorithm. (c) Gibbs stereo image after 140 iterations.

external field, we needed a more informed ordering. In particular, we ordered variables by their *restricted entropy* (i.e. the entropy of their distribution under only their external potential), then started with the most constrained variable and expanded via connectedness using entropy to break ties. This is one reasonable extension of the "most constrained first" approach to variable choice in deterministic constraint satisfaction. The quality of approximate samples with very few particles is encouraging, suggesting that appropriate sequentialization with arc consistency can leverage strong correlations to effectively move through the sample space.

4 Discussion

Our experiments suggest that tools from systematic search, appropriately generalized, can mitigate problems of multimodality and strong correlation in sampler design. When variables (and their attendant soft constraints) are incorporated one at a time, a sampler may be able to effectively find high probability regions by managing correlations one variable at a time. Additionally, any sample produced by sequential rejection is, by definition, exact.

Prior work generating samples from very large (and sometimes critical) ferromagnetic Isings has instead relied on specialized cluster sampling methods to manage the correlation problem. Given a *convergent* Markov chain, coupling from the past techniques provide termination analysis; i.e., they give a proof of exact convergence. Comparing and combining these approaches seems fruitful, and would likely build on the theory of sequential Monte Carlo samplers (and in particular, backward kernels) from (Del Moral et al., 2006) and (Hamze and de Freitas, 2005). In the case of approximate sampling, other work has used deterministic search as the subroutine of a sampler (Gogate and Dechter, 2007), rather than recovering search behavior in a limit. (Southey et al., 2002) uses local (not systematic) search to improve the quality of the proposal distribution for an importance sampler.

The rejection rate plots for Ising models show that our algorithm runs into difficulty near the phase transition, where the distribution is the most complex. Its effectiveness may track semantic features of the distribution, and it would be interesting to study this relationship analytically. It would also be interesting to explore SAT problems, which are known to empirically exhibit a phase transition in hardness. It would also be interesting to see how the rejection rate and memory consumption of adaptation in our algorithm relate to the cost of dynamic programming (ala junction tree), and to explore the behavior of straightforward blocked variants of our method where multiple variables are added simultaneously.

Exact sampling may be truly intractable for large problems, with exact samplers useful primarily as a source of proposals for approximate algorithms. However, it seems that recursive control structures from combinatorial optimization may be generally useful in sampler design, and encourage the development of samplers whose efficiency actually improves as soft constraints harden and probabilistic reasoning problems turn into satisfiability. By constraining our sampling algorithms to sample uniformly from satisfying solutions in the deterministic limit, we may arrive at more useful methods for the uncertain reasoning problems central to probabilistic AI.

Acknowledgements

This work was partially supported by gifts from NTT Communication Science Laboratories, the Eli Lilly Corporation and Google, and National Science Foundation fellowships. The authors would like to thank Nando de Freitas and Kevin Murphy for helpful discussions, Keith Bonawitz for clarifying the notion of restrictions, and Jonathan Gonda for accidentally providing critical inspiration.

References

A. M. Childs, R. B. Patterson, and D. J. C. MacKay. Exact sampling from non-attractive distributions using summary states. *arXiv:cond-mat/0005132v1*, 2000.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte

Carlo Samplers. *Journal of the Royal Statistical Society*, 68(3):411–436, 2006.

R. G. Edwards and A. D. Sokal. Generalizations of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review*, 38:2009–2012, 1988.

V. Gogate and R. Dechter. Samplesearch: A scheme that searches for consistent samples. In *Artificial Intelligence and Statistics*, 2007.

F. Hamze and N. de Freitas. Hot coupling: A particle approach to inference and normalization on pairwise undirected graphs. In *Advances in Neural Information Processing Systems 17*, 2005.

M. Huber. A bounding chain for Swendsen–Wang. *Random Structures & Algorithms*, 22(1):53–59, 2002.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2): 223–252, 1996.

L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(4):61–76, 1996.

D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1393–1400. MIT Press, 2008.

F. Southey, D. Schuurmans, and A. Ghodsi. Regularized greedy importance sampling. In *Advances in Neural Information Processing Systems 14*, pages 753–760, 2002.

R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in monte carlo simulations. *Physics Review Letters*, 58(2):86–88, Jan 1987. doi: 10.1103/PhysRevLett.58.86.

M. F. Tappen and W. T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Int. Conf. on Computer Vision*, 2003.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. In *Uncertainty in Artificial Intelligence*, pages 536–543, 2002.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems 13*, pages 689–695. MIT Press, 2001.

5 Appendix

If the distribution $P(\mathbf{x})$ is not the target distribution but instead the distribution at some intermediate stage in a sequential rejection sampler, the downstream stage will adapt P to match *its* marginal. Let $R \subset \mathbf{X}$, and consider the adapted distribution $P_R(\mathbf{x})$ with additional factors $\phi_{\mathbf{x}} \in [0,1]$ for $\mathbf{x} \in R$. For $\mathbf{x} \notin R$, let $\phi_{\mathbf{x}} = 1$. We show that these additional factors satisfy the existing pre-computed bound C^* and that sequential rejection on the adapted distribution P_R eventually accepts every sample. In this case, the weight of a sample is

$$W_{P' \to P_R}(\mathbf{y}, z) \triangleq \frac{\bar{P}_R(\mathbf{y}, z)}{\bar{P}'(\mathbf{y})\, Q_{P_R}(z \mid \mathbf{y})} \tag{24}$$

$$= \sum_{z'} \left(\psi_2(\mathbf{y}, z') \prod_{\mathbf{x} \in R} \phi_{\mathbf{x}}^{\delta_{\mathbf{x}(\mathbf{y},z')}} \right) \tag{25}$$

and therefore

$$W_{P' \to P_R}(\mathbf{y}, z) \leq \sum_{z'} \psi_2(\mathbf{y}, z') = W_{P' \to P}(\mathbf{y}). \tag{26}$$

We claim that $W_{P'_S \to P_R}(\mathbf{y}, z) \leq C^*_{P' \to P}$. Let R' and ϕ' be the set of $\mathbf{x} = (\mathbf{y}, z)$ and weights that have been fed back to P in previous iterations of the algorithm. Consider

$$W_{P'_S \to P_R}(\mathbf{y}, z) \triangleq \frac{\bar{P}_R(\mathbf{y}, z)}{\bar{P}'_S(\mathbf{y})\, Q_{P_R}(z \mid \mathbf{y})} \tag{27}$$

$$= \frac{W_{P' \to P_R}(\mathbf{y})}{\prod_{\mathbf{y}' \in S} \left(\frac{W_{P' \to P_{R'}}(\mathbf{y}')}{C^*_{P' \to P}} \right)^{\delta_{\mathbf{y}\mathbf{y}'}}} \tag{28}$$

$$= \begin{cases} W_{P' \to P_R}(\mathbf{y}) & \mathbf{y} \notin S \\ \frac{W_{P' \to P_R}(\mathbf{y})}{W_{P' \to P_{R'}}(\mathbf{y})} C^*_{P' \to P} & \mathbf{y} \in S. \end{cases} \tag{29}$$

Eq. (26) implies that when $\mathbf{y} \notin S$, we have $W_{P'_S \to P_R}(\mathbf{y}, z) \leq W_{P' \to P}(\mathbf{y}) \leq C^*_{P' \to P}$. Therefore, the claim is established for $\mathbf{y} \in S$ if $\frac{W_{P' \to P_R}(\mathbf{y})}{W_{P' \to P_{R'}}(\mathbf{y})} \leq 1$. We have that

$$\frac{W_{P' \to P_R}(\mathbf{y})}{W_{P' \to P_{R'}}(\mathbf{y})} = \frac{\sum_{z'} \psi_2(\mathbf{y}, z') \prod_{\mathbf{x} \in R} \phi_{\mathbf{x}}{}^{\delta_{\mathbf{x}(\mathbf{y},z')}}}{\sum_{z'} \psi_2(\mathbf{y}, z') \prod_{\mathbf{x} \in R'} \phi'_{\mathbf{x}}{}^{\delta_{\mathbf{x}(\mathbf{y},z')}}} \tag{30}$$

First note that, $\mathbf{x} \in R' \implies \mathbf{x} \in R$. Therefore, the inequality is satisfied if $\phi'_{\mathbf{x}} \geq \phi_{\mathbf{x}}$ for all $\mathbf{x}$. We prove this inductively. When a value $\mathbf{x}$ is first added to R, $\mathbf{x} \notin R'$, hence $\phi'_{\mathbf{x}} = 1 \geq \phi_{\mathbf{x}}$. By induction, we assume the hypothesis for $\phi_{\mathbf{x}}$ and show that $\phi'_{\mathbf{y}} \geq \phi_{\mathbf{y}}$. Consider Eq. (29). If $\mathbf{y} \notin S$, then $\phi_{\mathbf{y}} = \frac{W_{P' \to P_R}(\mathbf{y})}{C^*_{P' \to P}} \leq \frac{W_{P' \to P}(\mathbf{y})}{C^*_{P' \to P}} \leq 1 = \phi'_{\mathbf{y}}$ by the optimality of C^* and Eq. 26. If $\mathbf{y} \in S$, we have $\phi'_{\mathbf{x}} \geq \phi_{\mathbf{x}}$ for all $\mathbf{x}$ by induction, proving the claim.

Evidently, the weights decrease monotonically over the course of the algorithm. Of particular note is the case when $R = \mathbf{X}$ and $S = \mathbf{Y}$: here the acceptance ratio is again 1 and we generate exact samples from P_R. Of course, $|R|$ and $|S|$ are bounded by the number of iterations of the algorithm and therefore we expect saturation (i.e., $|R| = |\mathbf{X}|$) only after exponential work.