Dynamic Scaled Sampling for Deterministic Constraints

Lei Li
University of California, Berkeley
leili@cs.berkeley.edu

Bharath Ramsundar
Stanford University
rbharath@stanford.edu

Stuart Russell
University of California, Berkeley
russell@cs.berkeley.edu

Abstract

Deterministic and near-deterministic relationships among subsets of random variables in multivariate systems are known to cause serious problems for Monte Carlo algorithms. We examine the case in which the relationship $Z = f(X_1, \ldots, X_k)$ holds, where each X_i has a continuous prior pdf and we wish to obtain samples from the conditional distribution $P(X_1, \ldots, X_k \mid Z = s)$. When f is addition, the problem is NP-hard even when the X_i are independent. In more restricted cases—for example, i.i.d. Boolean or categorical X_i—efficient exact samplers have been obtained previously. For the general continuous case, we propose a *dynamic scaling* algorithm (DYSC), and prove that it has $O(k)$ expected running time and finite variance. We discuss generalizations of DYSC to functions f described by binary operation trees. We evaluate the algorithm on several examples.

1 Introduction

Monte Carlo methods for inference are among the most useful tools in machine learning and computational statistics models. For example, posterior marginals in Bayesian networks can be approximated by rejection sampling [15], importance sampling [12, 23], or Markov chain Monte Carlo [20].

For these standard sampling algorithms, deterministic or near-deterministic relationships—which are ubiquitous in real-world applications (e.g., models from physics, chemistry, systems biology, economics, computer vision) —cause serious problems. In MCMC, the problem occurs regardless of whether the constraint variables are observed, for their values will be set by sampling. Standard convergence proofs fail with deterministic variables, while near-deterministic relationships cause approximation bounds to become arbitrarily large and MCMC mixing rates to become arbitrarily slow. These difficulties were noted by Chin and Cooper [5] among others; their proposed solutions involve transforming the network to eliminate the offending variable. Unfortunately, such transformations can render the network intractably large and can themselves be computationally infeasible to carry out. Thus, the problem remains important in practice.

In this paper, we take a different approach by developing local samplers that remain efficient despite the presence of deterministic constraints. The core problem we consider is shown in Fig. 1(a): a variable Z depends deterministically on parents $X_1, \ldots, X_k$ via $Z = f(X_1, \ldots, X_k)$. Independent priors $P_i(x_i)$ and the fixed value $Z = s$ are given, and we wish to sample from the posterior $P(x_1, \ldots, x_k \mid Z = s)$. As well as being of independent interest, such a local sampler can be used as a component within an overall MCMC architecture to resample the entire parent set of a deterministic variable or to sample any subset of its parents given values for the others. In the more general case (Fig. 1(b)), the parent variables may be dependent and the entire fragment may be embedded within a larger network.

For most of the paper, we focus on the case where the deterministic function f is addition, i.e., we wish to sample a set of variables conditioned on their sum:

Problem 1 (Sampling from posterior given sum)**.** *Given s, and k* independent *random variables $X_1, \ldots, X_k$, such that $X_i \sim P_i(x)$, sample from the posterior distribution $P(x_1, \ldots, x_k \mid \sum_{i=1}^{k} x_i = s)$.*

This simple problem arises in many applications, including unusual event discovery [19], system component failure detection [16], estimating Internet topology [10, 18], multitarget tracking [24, 28], sybil attack [9], and image segmentation [25].

We describe relevant prior work in Sec. 2, including

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

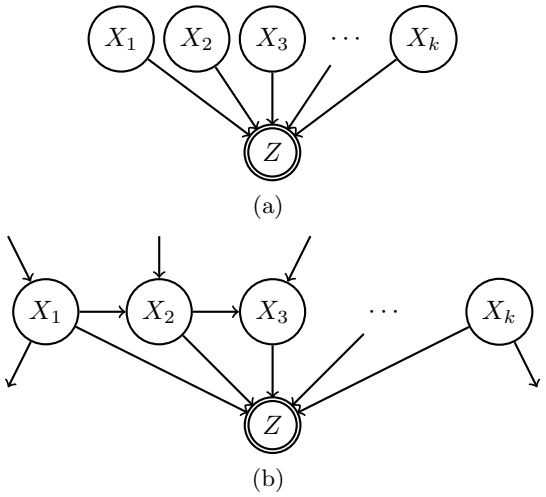


Figure 1: (a) A simple Bayesian network with one observed child that depends deterministically on k independent parents. (b) A more general setting with dependent parents and additional network context.

some known special cases that are efficiently solved; for example, the binary case can be solved using dynamic programming [4] or belief propagation with FFT [26]. We show in Sec. 3 that the general cases for both discrete and continuous variables are NP-hard. Sec. 4 describes DYSC, an importance sampling algorithm for the general case; we prove that DYSC generates samples satisfying the deterministic constraint in time linear in k, and show that it works well in practice. Sec. 6 shows that the same algorithmic scheme may be applied to any deterministic function that is described by a binary operation tree with computable inverses for its nodes.

2 Background and Previous Approaches

Tierney [27] summarizes the MCMC approach to sampling from posterior distributions; the basic convergence theory requires connectedness of the state space, but, with deterministic constraints on variables, the required conditions may be absent when variables are sampled individually. Connectedness may be restored by sampling subsets of variables jointly, which is exactly the problem studied in this paper.

Importance sampling and its resampling variant SIR [21] are often used as an alternative to MCMC. They rely heavily on the choice of a "good" proposal distribution; moreover, with deterministic evidence on continuous variables (such as the given value s in Problem 1), importance sampling can degenrate to rejection sampling with a 100% probability of rejection. Fung and del Favero [13] propose a backward simulation method that can sample from the posterior distribution given evidence; its efficiency relies, however, on a complete tabulation of the conditional distributions—in our case, $P(Z \mid X_1, \ldots, X_k)$—which would be exponentially large for discrete variables and impossible for continuous variables.

Gogate and Dechter [14] handle sampling with deterministic evidence with a backtracking search algorithm, i.e. through searching for samples consistent with the given evidence constraints. However, their algorithm does not work for continuous variables.

For the case of summation evidence, there have been several solutions that assume special forms for the prior distributions. In particular, one can sample exactly from the posterior distribution on a set of k Bernoulli variables, given their sum s, using dynamic programming [16]. The complexity is $O(s^2 + sk)$, or $O(sk)$ if weights can be pre-computed. The algorithm is the same as the maximum entropy method [4]. When $s < k$, the posterior may be computed even more efficiently ($O(k \log^2 k)$) by using the sum–product algorithm with FFT [26]. This FFT method can be extended to categorical random variables [11]. However, the same approach does not apply to continuous variables with general densities.

For discrete variables, contingency table sampling provides an alternative approach to sample from posterior distributions [7, 8]. For i.i.d. Bernoulli variables, the problem reduces to that of choosing s of k variables uniformly without replacement [2].

The rare-event literature [3, 1] considers the problem of accurately estimating the probabilities of events with low, but nonzero probability. These works are philosophically related to ours. However, it is not clear how to apply algorithms from this literature to solve Problem 1, because the constrained distributions we consider have zero prior probability.

3 How Hard is the Problem?

Before describing our sampling algorithm, we first analyze the hardness of our inference problem. It is straightforward to show that Problem 1 is NP-hard in general. The proof follows from the NP-hardness of the restricted additive constraint problem where each variable has a categorical distribution over integer values. We consequently prove hardness results for two refined versions of Problem 1. The first version proves that drawing samples from the posterior distribution is NP-hard when the X_i are discrete. The second version proves that drawing high quality samples from the posterior distribution is NP-hard when the X_i are

real-valued.

Problem 2 (Posterior for integer-valued variables)**.** *Given s, and k* independent *random variables $X_1, \ldots, X_k$ with finite support over the* non-negative integers *such that $P_i(X_i = j) = p_i^j$ (p_i^j can be represented using finite bits), the goal is to generate samples $P(X_1, \ldots, X_k \mid \sum_{i=1}^k X_i = s)$ with posterior probability greater than zero.*

Theorem 1. *Problem 2 is NP-hard.*

Proof. The proof is by reduction from the SUBSET-SUM problem, which is known to be NP-complete [6]. An instance of the SUBSET-SUM problem is to select a subset A from $\{a_1, a_2, \ldots, a_k\}$ such that the sum of the set is s, where a_i are positive integers. We construct an corresponding instance of Problem 2 as follows: $P_i(x_i = a_i) = 1/2$, $P_i(x_i = 0) = 1/2$. We only need $\log a_i$ bits and one additional bit to represent P_i. Note we do not represent any values with zero probability. Hence the reduced problem has a polynomial size compared to the original SUBSET-SUM problem. Clearly, if we could obtain any consistent sample, say $\hat{x_1}, \hat{x_2}, \ldots \hat{x_k}$ such that $\hat{x_1} + \hat{x_2} + \cdots + \hat{x_k} = s$, we could then get a solution to the original SUBSET-SUM problem $A = \{\hat{x_i} : \hat{x_i} > 0\}$. SUBSET-SUM is NP-complete, therefore Problem 2 is NP-hard. □

In contrast, for variables with continuous support, it is often easy to generate joint samples that satisfy the sum constraint. On the other hand, sampling from the *posterior distribution*, rather than just producing any old values satisfying the sum constraint, may be harder. The following is a definition of the problem of generating "good" samples.

Problem 3 (Posterior with bounded non-negative continuous variables)**.** *Let s be a target sum, ϵ be a bounded-precision positive number, and $X_1, \ldots, X_k$ be independent real-valued random variables, such that $X_i \sim p_i(x)$. Assume that $p_i(x) > 0$ for $x \in [0, \infty)$ and that $p_i(\cdot)$ is defined by finitely many bounded-precision parameters. The goal is to find a sample (x_i) from the posterior probability density $p(X_1, \ldots, X_k \mid \sum_{i=1}^k X_i = s)$ such that $\sum_{i=1}^k X_i = s$ and the un-normalized probability density of the sample satisfies*

$$p(X_1 = x_1) \cdot p(X_2 = x_2) \cdot \cdots \cdot p(X_k = x_k) > \epsilon \ . \quad (1)$$

If such an (x_i) does not exist, failure must be reported.

Theorem 2. *Problem 3 is NP-hard.*

Proof sketch (full details in the supplementary material): The proof is again by reduction from SUBSET-SUM. We start with an instance in SUBSET-SUM, with a set of nonzero integers $\{a_1, \ldots, a_k\}$ and the target sum s. The strategy will be to define p_i to be a continuous bimodal density with almost all mass close to a_i and 0, such that any valid solution to SUBSET-SUM leads to a solution to the sampling problem with prior p_i, and vice versa. The prior densities are carefully chosen with compact representation in terms of bits:

$$p_i(x) = \frac{I(x \geq 0)}{c_i} \left(\frac{1}{2}\mathcal{N}(0, \delta^2) + \frac{1}{2}\mathcal{N}(a_i, \delta^2) \right)$$

where I is the indicator function and $\mathcal{N}(\cdot, \cdot)$ is the normal distribution density and $\delta = \frac{1}{8k}$. c_i is a normalizing constant given by $\frac{1}{4} + \frac{1}{2}\phi(-\frac{a_i}{\delta})$, where ϕ is CDF of the standard normal distribution. Let $c = p(\sum_{i=1}^k X_i = s) > 0$, and we set

$$\epsilon = \frac{1}{\prod_i c_i} \cdot \frac{1}{(2\delta\sqrt{2\pi})^k} \prod_i \left(1 + \exp(-\frac{a_i^2}{\delta^2}) \right) \cdot \eta$$

where $\eta = 2e^{-16} < 1$. Here we assume that real values $\delta, \epsilon, \eta, c_i$ can be represented with a constant number of bits.

Suppose that there exists such a subset $A \subseteq \{a_1, \ldots, a_k\}$ satisfying $\sum_{a \in A} a = s$. This subset will certainly lead to a valid sample, i.e., $X_i = a_i$ if $a_i \in A$ and $X_i = 0$ otherwise. It is then straightforward to verify that the sample satisfies Eq. (1).

Conversely, suppose there is a sample $x_1, \ldots, x_k$ satisfying Eq. (1) and $\sum_i x_i = s$. We claim $|x_i| < \frac{1}{2k}$ or $|x_i - a_i| < \frac{1}{2k}$ for each $i = 1 \ldots k$. We can select the set $A = \{a_i \mid |x_i - a_i| < \frac{1}{2k}\}$ with sum s.

Easy cases In some special cases, there exist fast algorithms to sample exactly from the posterior distribution.

Example 1. *k i.i.d. binary rv with target sum s. The posterior is uniform among $\binom{s}{k}$ configurations [16] which can be generated in polynomial time.*

Example 2. *k i.i.d. rv from exponential priors, with the target sum s. The posterior is uniformly distributed in the $k-1$ simplex. There is a simple and fast algorithm (Alg. 3.23 in [17]) to generate such samples: first draw k values independently from the prior exponential distribution, and then normalize accordingly.*

4 DYSC: Proposed Method for Continuous Variables

For Problem 1 with continuous variables, we propose a sequential importance sampling algorithm called DYSC (DYnamic SCaling). For simplicity of exposition, this section assumes that the X_is are i.i.d. with prior density $p(\cdot \mid \theta)$, where θ is the mean. These restrictions are not essential.

The basic intuition for the algorithm can be obtained by considering the failure modes of a naive method, i.e., sampling from the prior for $X_1, \ldots, X_{k-1}$ and then setting X_k to the value required such that the sum is s. Clearly, if s is much larger than $k\theta$, X_k will probably have a very large value and samples generated in this way will be highly asymmetric and have mostly very small weights. Conversely, if s is small, the sum of the first few values may exceed s and the sample must be rejected (or have all the remaining values set to 0). DYSC tries to avoid these problems by dynamically scaling a parameterized proposal distribution $q_i(\cdot|\eta_i)$ for each X_i so that its mean η_i equals the amount needed per remaining variable such that the desired total s is reached.

The proposal distributions q_i are designed in one-to-one correspondence with the terms in the factorized form of the posterior distribution:

$$P(x_1, \ldots, x_k \mid \sum_{i=1}^{k} x_i = s) = \prod_{i=1}^{k} P(x_i \mid x_1, \ldots, x_{i-1}, \sum_{j=i}^{k} x_j = s - \sum_{k=1}^{i-1} x_i) \quad (2)$$

We set $q(X_1, \ldots, X_k|s) = q_1(X_1|s) \cdot q_2(X_2|s, x_1) \cdots q_k(X_k|s, x_1, \ldots, x_{k-1})$.

The proposal distributions q_i are subject only to the usual conditions for importance sampling; but the simplest idea is to rescale the prior to have the desired mean, so we can set q_i to be a suitably truncated version of $p(\cdot \mid \eta_i)$:

$$\eta_i = \frac{s - \sum_{j=1}^{i-1} x_j}{k - i + 1} \quad (3)$$

$$q_i(X_i|\eta_i) = \begin{cases} \frac{p(X_i|\eta_i) I(0 \le X_i \le S - \sum_{j=1}^{i-1} x_j)}{\int_0^{S - \sum_{j=1}^{i-1} x_j} p(y|\eta_i) dy} & \text{if } \sum_{j=1}^{i-1} x_j < s \\ \delta(0) & \text{otherwise} \end{cases} \quad (4)$$

$$q_k(X_k) = \delta(X_k = s - x_1 - \cdots - x_{k-1}) \quad (5)$$

where $I(\cdot)$ is the indicator function and $p(\cdot \mid \eta_i)$ is the prior of X_i with parameter η_i. Therefore, the importance weight can be calculated as

$$w = \frac{p(x_1, \ldots, x_k, S)}{q(x_1, \ldots, x_k, S)} = p(S - \sum_{j=1}^{k-1} x_j \mid \theta) \cdot \prod_{i=1}^{k-1} \left(\frac{p(x_i \mid \theta)}{q_i(x_i|\eta_i)} \right) \quad (6)$$

The whole algorithm is described in Alg. 1; notice that truncation is achieved by rejection sampling in line 8 and the proposal $\tilde{q}$ is non-truncated. The following sections develop a per-sample runtime analysis.

Algorithm 1: DYSC: sampling posterior for continuous variables

Input: s: the target sum; k: the number of variables; $X_i \sim p(\cdot|\theta)$; $\tilde{q}(\cdot|\eta)$: a family of importance distributions parameterized by mean η, with $c(\cdot \mid \eta)$ its cumulative function;

Output: A sample for $X_1, \ldots, X_k$ with weight w

1 initialize all $X_i \leftarrow 0$, $R \leftarrow s$, $w \leftarrow 1$;
2 **for** $i \leftarrow 1$ *to* $k-1$ **do**
3 **if** $R = 0$ **then**
4 $X_i \leftarrow 0$; $\quad w \leftarrow w \cdot p(0|\theta)$;
5 **else**
6 $\eta_i \leftarrow \frac{R}{k-i+1}$;
7 construct proposal $\tilde{q}_i(\cdot) = \tilde{q}(\cdot|\eta_i)$;
8 **repeat** sample $X_i \sim \tilde{q}_i(\cdot)$ **until** $X_i \in [0, R]$;
9 $w \leftarrow w \cdot (c_i(R|\eta_i) - c_i(0|\eta_i)) \cdot \frac{p(X_i|\theta)}{\tilde{q}(X_i)}$;
10 $R \leftarrow R - X_i$;
11 $X_k \leftarrow R$; $\quad w \leftarrow w \cdot p(R)$;

4.1 Basic assumptions

For the purposes of this analysis, we will make the following assumptions.

1. The random variables X_i are independent and identically distributed.
2. X_i has continuous probability density function p_i with respect to Lebesgue measure.
3. Importance distributions $q_i(\cdot|\eta_i)$ are nonzero continuous probability density functions on $[0, \infty)$. These distributions vary continuously in parameter η_i. q_i has a positive continuous lower bound function f_i that does not depend on choice of η_i.

The independence assumption in (1) is not strictly required but gives a nice form to the importance weights. Assumption (2) could be weakened slightly, but doing so would complicate our proofs. Assumption (3) has some subtleties. Let q denote the distribution that DYSC actually samples from (the aggregate of the q_i). For DYSC to have finite variance, we will need q to be lower bounded regardless of the η_i chosen. Assumption (3) forces this lower bound. Note on the other hand that q need not be upper bounded. Consider the exponential distribution $f(x; \lambda) = \lambda e^{-\lambda x}$, where $\mu = \frac{1}{\lambda}$ is the mean. As $\mu \to 0$, $\lambda \to \infty$, so f grows unboundedly near $x = 0$ with shrinking μ. If q_i has distribution f, then q may grow unboundedly as well.

4.2 Geometry of the simplex

We begin our analysis by building some geometric intuition about the space we seek to sample from. For brevity, we only provide proof sketches in the sequel. Complete proofs may be found in the supplementary material.

Definition 1. *Let x be any positive real number. A standard k-simplex of value x is a subset of $\mathbb{R}^{k+1}$ given by*

$$\Delta_x^k = \{(t_0, \cdots, t_k) \in (\mathbb{R})^{k+1} \mid \sum_{i=0}^{k} t_i = x,\ \forall i\ t_i \geq 0\}$$

We will use the following two facts about the geometry of Δ_S^{k-1} to study Alg. 1.

Lemma 1. *Δ_x^k is compact.*

Lemma 2. *Let f be a continuous real-valued function defined on compact set X. Then f is bounded on X.*

Recall that the support of a random variable indicates the set of values it can assume with positive probability. When defining the support for real-valued random variables with continuous density, the following definition suffices

Definition 2. *Let X be a random variable taking values in $\mathbb{R}^n$ with continuous pdf f.*

$$\text{supp}(X) = \{e \in \mathbb{R}^n \mid f(e) > 0\}$$

Lemma 3. *The target joint distribution $X_1, \ldots, X_k, \sum_{i=1}^k X_i = S$ has continuous probability density function p on R^k with support containing the standard $k-1$ simplex of value S. That is,*

$$\text{supp}(X_1, \ldots, X_k, \sum_{i=1}^{k} X_i = S) \supseteq \Delta_S^{k-1}\ .$$

Let $\text{DYSC}(S, k, X_i, q_i)$ denote the $(x_1, \ldots, x_k)$ sampled in $\mathbb{R}^k$ by Alg. 1. We analyze the behavior of this random variable. We start by using Markov's inequality to prove an important property of the dynamic scaling.

Lemma 4. *If $S - \sum_{j=1}^{i-1} x_j > 0$ then*

$$P_{q_i(\cdot|\eta_i)}(0 \leq X_i \leq S - \sum_{j=1}^{i-1} x_j) \geq \frac{1}{2}$$

Proof sketch: The bound can be derived from the equation

$$\mathbb{E}_{q_i(\cdot|\eta_i)}[X_i] = \int x_i q_i(x_i|\eta_i) dx_i = \eta_i = \frac{S - \sum_{j=1}^{i-1} x_j}{k - i + 1}$$

Lemma 5. *$\text{DYSC}(S, k, X_i, q_i)$ has nonzero probability density function q on $\mathbb{R}^k$ with support equal to the standard $k-1$ simplex of value S*

$$\text{supp}(\text{DYSC}(S, k, X_i, q_i)) = \Delta_S^{k-1}$$

q is continuous in the interior of Δ_S^{k-1}, $\text{int}(\Delta_S^{k-1})$, and is lower bounded on Δ_S^{k-1} by some $c > 0$.

Proof sketch: We prove this result by constructing an explicit formula for q on $\mathbb{R}^k$, and by exploiting Assumption (3).

Lemma 6. *Let $W(x) = \frac{p(x)}{q(x)}$. Then W is a bounded function on Δ_S^{k-1} that is continuous on $\text{int}(\Delta_S^{k-1})$*

Proof. From lemma 3, $p(x)$ is positive and continuous on Δ_S^{k-1}, and from lemma 5, $q(x)$ is lower bounded on Δ_S^{k-1} by $c > 0$ and continuous on $\text{int}(\Delta_S^{k-1})$. Hence W is nonzero and bounded on Δ_S^{k-1} and continuous on $\text{int}(\Delta_S^{k-1})$. □

4.3 Correctness Proofs

Theorem 3. *Let E be any measurable subset of the $k-1$ simplex. Let p_E denote the probability that the true joint takes values in E. Suppose we draw N times from variable $\text{DYSC}(S, k, X_i, q_i)$ and obtain $Y_1, \ldots, Y_N$ taking values in Δ_S^{k-1}. Then the following equation provides an unbiased estimator of p_E:*

$$\widehat{p_E} = \frac{1}{N} \sum_{i=1}^{N} I(Y_i \in E) W(Y_i)$$

Proof sketch: The proof is the standard correctness proof for importance samplers.

$$\begin{aligned}\mathbb{E}_q[\widehat{p_E}] &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_q[I(Y_i \in E) W(Y_i)] \\ &= \frac{1}{N} \sum_{i=1}^{N} \int q(y) I(y \in E) \frac{p(y)}{q(y)} dy = p_E\end{aligned}$$

Theorem 4. *$\widehat{p_E}$ has finite variance that shrinks to 0 as the number of samples N goes to infinity.*

Proof sketch: Recall that the Y_i are independent samples from $\text{DYSC}(S, k, X_i, q_i)$

$$\begin{aligned}\mathbb{V}\text{ar}_q[\widehat{p_E}] &= \frac{1}{N^2} \sum_{i=1}^{n} \mathbb{V}\text{ar}_q[I(Y_i \in E) W(Y_i)] \\ &= \frac{1}{N^2} \sum_{i=1}^{n} \left(\int p(y) I(y \in E) W(y) dy - p_E^2 \right)\end{aligned}$$

To simplify this equation note that by lemmas 6 and 3, W and p are bounded on Δ_S^{k-1}. Then there exists $C > 0$ that upper bounds $p(x)W(x)$ on Δ_S^{k-1}. Hence

$$\int p(y)I(y \in E)W(y)dy < C\mu(E) < \infty$$

where μ is the Lebesgue measure. Thus we can write

$$\mathbb{V}\mathrm{ar}_q[\widehat{p_E}] = \frac{1}{N^2}\sum_{i=1}^{n}\left(\int p(y)I(y \in E)W(y)dy - p_E^2\right)$$
$$< \frac{1}{N^2}\sum_{i=1}^{N}\left(C\mu(E) - p_E^2\right) = \frac{C\mu(E) - p_E^2}{N} \to 0$$

4.4 Runtime Bounds

We now analyze the expected runtime of Alg. 1. We start by recalling some basic facts about Geometric random variables.

Definition 3. Geom(p) *denotes a geometric random variable and counts the number of tails seen before heads in a sequence of* Bernoulli(p) *coin tosses. The following are basic facts about* Geom(p)

$$\mathbb{E}[\mathrm{Geom}(p)] = \frac{1-p}{p} \quad \mathbb{V}\mathrm{ar}[\mathrm{Geom}(p)] = \frac{1-p}{p^2}$$

Since line 8 contains a rejection sampling step, the runtime is itself a random variable.

Definition 4. *Let* $R_{\mathrm{DYSC}}(S, k, X_i, q_i)$ *denote the total number of failed rejection sampling steps in a run of Alg. 1. Let* $T_{\mathrm{DYSC}}(S, k, X_i, q_i)$ *denote the total number of rejection sampling steps in a run of Alg. 1.*

Recall from lemma 4 that

$$P_{q_i(\cdot|\eta_i)}(0 \le X_i \le S - \sum_{j=1}^{i-1} x_j) \ge \frac{1}{2}$$

Consequently, the probability that a rejection sampling step will fail is upper bounded by $\frac{1}{2}$.

Lemma 7.

$$\mathbb{E}[R_{\mathrm{DYSC}}(S, k, X_i, q_i)] \le k - 1$$

Proof sketch: The number of failed rejection sampling steps in iteration i at line 8 is geometrically distributed with some probability $p_i \ge 1/2$. Thus, we obtain

$$\mathbb{E}[R_{\mathrm{DYSC}}(S, k, X_i, q_i)] = \sum_{i=1}^{k-1}\mathbb{E}[\mathrm{Geom}(p_i)]$$
$$\le \sum_{i=1}^{k-1}\mathbb{E}[\mathrm{Geom}(\frac{1}{2})] = \sum_{i=1}^{k-1}\frac{1-\frac{1}{2}}{\frac{1}{2}} = k - 1$$

Lemma 8.

$$\mathbb{V}\mathrm{ar}[R_{\mathrm{DYSC}}(S, k, X_i, q_i)] \le 2k - 2$$

Proof sketch:

$$\mathbb{V}\mathrm{ar}[R_{\mathrm{DYSC}}(S, k, X_i, q_i)] = \mathbb{V}\mathrm{ar}[\sum_{i=1}^{k-1}\mathrm{Geom}(p_i)]$$
$$= \sum_{i=1}^{k-1}\mathbb{V}\mathrm{ar}[\mathrm{Geom}(p_i)] \le \sum_{i=1}^{k-1}\mathbb{V}\mathrm{ar}[\mathrm{Geom}(\frac{1}{2})] = 2k - 2$$

Lemma 9.

$$P(T_{\mathrm{DYSC}}(S, k, X_i, q_i) \ge 4k - 3) \le \frac{1}{2k-2}$$

Proof sketch: Apply Chebyshev's inequality to lemmas 7 and 8.

5 Experiments

We have performed experiments on several cases to test both the efficiency and sample quality of our algorithm. In particular, we would like to analyze the effects of dynamic scaling of proposal distributions. To this end, the base method is importance sampling with the same proposal distributions as DYSC but without scaling (IS for short in the following). For IS, we employ for each variable X_i the same rejection-based method as DYSC.

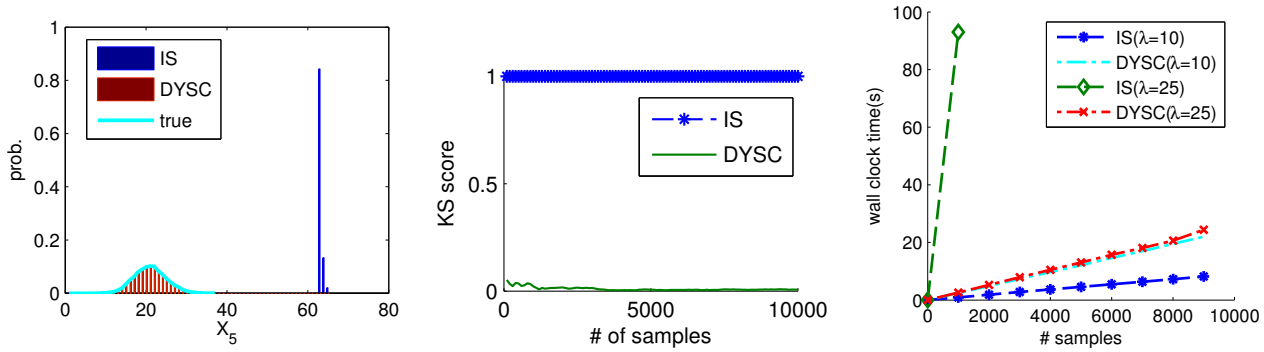
5.1 Effectiveness of DYSC

To evaluate the quality of DYSC, we set the prior distribution to be discrete in order to exactly calculate the true posterior distribution (The DYSC algorithm can be extended in a straightforward manner to discrete random variables). In our first experiment, we adopt the following example.

Case 1 k i.i.d. variables $X_i \sim p(X) = \mathrm{Poisson}(\lambda)$ such that $\sum_{i=1}^{k} X_i = s$.

We show the figures for the first example with various setting of k, s and λ. Fig 2(a) shows the marginal conditional distribution of X_5 for $k = 5$, $s = 100$, and $\lambda = 5$. The histograms are produced by generating 10,000 samples using each method.

Note the importance sampling without dynamic scaling of proposals produces empirical distribution far away from the true posterior. Fig. 2(b) shows the difference (measured by Kolmogorov–Smirnov score) of sampled x_5 from the true posterior as we draw more



(a) (b) (c)

Figure 2: (a): Histogram of 10,000 (weighted) samples for x_5 ($k = 5, s = 100$, $x_i \sim Poisson(\lambda)$ and $\lambda = 5$). Note importance sampling without dynamic scaling (IS) differs significantly from the true posterior. (b): Kolmogorov–Smirnov score for empirical distribution of x_5 against the true posterior, with the same setting as (a). Note DYSC quickly converges to zero (ideal) with more samples drawn. (c): Wall clock time for increasing number of samples ($k = 5$, $s = 50$, $\lambda = 10$ and $\lambda = 25$).

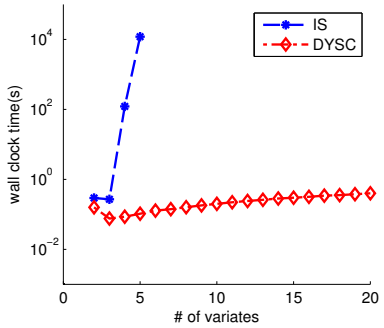


Figure 3: Wall clock time as a function of the number of variables ($s = 50$, $x_i \sim Poisson(\lambda)$, and $\lambda = 20$); 100 samples per run, averaged over 10 runs. Note the IS does not finish on large cases.

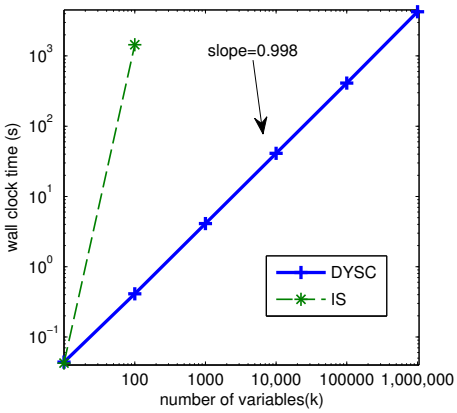


Figure 4: Wall clock time as a function of the number of variables ($s = 100$, $x_i \sim \text{LogNormal}(0, 1)$); 100 samples per run, averaged over 10 runs. Note the linear scalability of DYSC. In contrast, importance sampling without dynamic scaling (IS) can only finish when k is small.

samples ($s = 100$, $k = 5$ and $\lambda = 5$). Note DYSC quickly converges to zero (ideal) with more samples drawn while IS shows no clear sign of converging.

Fig. 2(c) shows the running time for $k = 5$, $s = 50$, $\lambda = 10$, and $\lambda = 25$. Note in both cases, DYSC has almost the same running time for different prior parameters, while IS varies drastically and does not finish in some cases. Fig. 5.1 shows the running time for varying number of variables ($s = 50$ and $\lambda = 20$). Note the IS method quickly consumes exponential time, while DYSC has linear time complexity in k.

5.2 Modeling phone calls

Mobile operators often want to predict consumer behavior by simulating the number of phone calls for each customer. It is observed that heavy-tailed distributions (e.g., LogNormal) model phone calls well in practice [22]. We will simulate numbers of phone calls for k customers with a given total sum.

Case 2 k i.i.d. variables $X_i \sim p(X) = \text{LogNormal}(\mu, \sigma)$ such that $\sum_{i=1}^{k} X_i = s$.

In our experiment, we test the running time of both importance sampling with and without dynamic scaling. DYSC can sample efficiently even with $k = 300,000$ and takes approximately 20 minutes to draw 100 samples in this case. DYSC scales linearly with increasing number of variables as shown in Fig. 4, while IS explodes when sampling more than ten variables.

6 Discussion and Generalization

Stability: DYSC can be viewed as a sequential importance sampler (SIS). In some cases, SIS methods diverge exponentially as the number of variables grows. DYSC modifies its proposal distribution in a forward looking way after each variable is sampled, so its behavior may be more robust than naive SIS on deterministic evidence.

Resampling: The success of sequential importance resampling (SIR) methods may suggest the addition of resampling during each step of DYSC. SIR methods are effective when we can calculate the conditional marginal distributions (e.g. $p(x_1|s)$, $p(x_2|x_1, s)$, ...). However, with deterministic evidence these conditional marginal distributions do not in general have closed form solutions.

Generalization: The approach of dynamically scaling proposal distributions can be extended to evidence involving other deterministic functions, e.g., weighted summation and multiplication, logical functions like AND, OR, XOR over Boolean values. In general, suppose the observed evidence is a function $Z = f(X_1, \ldots, X_k)$, where each variable has its own prior distribution. Again, the goal is to generate samples with respect to the posterior distribution of $P(X_1, \ldots, X_k | f(X_1, \ldots, X_k) = s)$. We can extend our DYSC algorithm to a class of functions that can be represented as binary operation trees with certain additional conditions.

When f is multiplication, a solution can be obtained using logarithms to reduce the problem to a summation constraint that DYSC can handle. When such tricks are not possible, a more general approach can be developed. We require that the "input variables" (variables with prior distribution) appear only once in the tree and that each binary operation be argument-wise invertible—that is, for each binary function $g_i(x, y)$ there should exist efficiently computable functions $h_i^1(z, y)$ and $h_i^2(z, x)$ such that $g_i(h_i^1(z, y), y) = z$ and $g_i(x, h_i^2(z, x)) = z$. Proofs of the runtime properties of DYSC on trees remain to be derived.

Another direction for future work is extending the algorithm to dependent random variables. The DYSC algorithm in this paper only considers cases with independent prior distributions. In the dependent cases (e.g., Fig. 1(b)), we can still decompose the expected sums into parts by linearity of expectation. We do not yet know how to scale the proposal properly to match the posterior expectation.

7 Conclusion

Standard sampling algorithms for Bayes nets often fail in the presence of deterministic constraints. In this paper, we propose a fast algorithm, DYSC, to handle such determinism, and demonstrate our solution for summation evidence over continuous random variables. We show that in general it is NP-hard to generate high quality samples from the posterior distribution conditioned on the target sum by reducing the corresponding decision problem to SUBSET-SUM. We discuss generalizations of our approach that handle a larger class of deterministic functions defined by binary operation trees. Our algorithms can serve as building blocks for the general inference problem in Bayesian networks; we envisage MCMC systems with large libraries of such "expert" sub-model samplers.

References

[1] Z. I. Botev and D. P. Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability*, 10(4):471–505, 2008.

[2] G. Broström and L. Nilsson. Acceptance-rejection Sampling from the Conditional Distribution of Independent Discrete Random Variables, given their Sum. *Statistics*, 34(3):247–257, Jan. 2000.

[3] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808, 2012.

[4] X. H. Chen, A. P. Dempster, and J. S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3), 1994.

[5] H. Chin and G. Cooper. Bayesian belief network inference using simulation. In *Uncertainty in Artificial Intelligence 3 Annual Conference on Uncertainty in Artificial Intelligence (UAI-87)*, pages 129–147, Amsterdam, NL, 1987. Elsevier Science.

[6] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[7] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1), 1998.

[8] A. Dobra, C. Tebaldi, and M. West. Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference*, 136(2):355–372, 2006.

[9] J. R. Douceur. The sybil attack. In *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, IPTPS '01, pages 251–260, London, UK, UK, 2002. Springer-Verlag.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99*, pages 251–262, New York, NY, USA, 1999. ACM.

[11] P. Felzenszwalb, D. Huttenlocher, and J. Kleinberg. Fast algorithms for large state space HMMs with applications to web usage analysis. In *Advances in Neural Information Processing Systems*, 2003.

[12] R. Fung and K.-C. Chang. Weighing and integrating evidence for stochastic simulation in Bayesian networks. In *Proceedings of the Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-89)*, pages 112–117, New York, NY, 1989. Elsevier Science.

[13] R. M. Fung and B. D. Favero. Backward simulation in Bayesian networks. In *UAI*, pages 227–234, 1994.

[14] V. Gogate and R. Dechter. Samplesearch: Importance sampling in presence of determinism. *Artif. Intell.*, 175(2):694–729, Feb. 2011.

[15] M. Henrion. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In *UAI*, pages 149–164, 1986.

[16] A. B. Huseby, M. Naustdal, and I. D. Varli. System reliability evaluation using conditional Monte Carlo methods. *in Statistical Res. Rep*, 2:0806–3842, 2004.

[17] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. Wiley, 2011.

[18] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.

[19] I. B. Mugtussids. *Flight Data Processing Techniques to Identify Unusual Events*. PhD thesis, Virginia Tech, 2000.

[20] J. Pearl. Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence*, 32:247–257, 1987.

[21] D. B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):pp. 543–546, 1987.

[22] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *KDD '08*, pages 596–604, New York, NY, USA, 2008. ACM.

[23] R. D. Shachter and M. A. Peot. Simulation approaches to general probabilistic inference on belief networks. In M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer, editors, *UAI*, pages 221–234. North-Holland, 1989.

[24] R. W. Sittler. An optimal data association problem in surveillance theory. *IEEE Transactions on Military Electronics*, 8(2):125 –139, 1964.

[25] E. B. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent Pitman–Yor processes. In *NIPS*, pages 1585–1592, 2008.

[26] D. Tarlow, K. Swersky, R. S. Zemel, R. P. Adams, and B. J. Frey. Fast exact inference for recursive cardinality models. In N. de Freitas and K. P. Murphy, editors, *UAI*, pages 825–834. AUAI Press, 2012.

[27] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

[28] J. Vermaak, S. J. Godsill, and P. Perez. Monte Carlo filtering for multi-target tracking and data association. *IEEE Transactions on Aerospace and Electronic Systems*, 41:309–332, 2005.