Symbolic Gibbs Sampling in Piecewise Algebraic Graphical Models with Nonlinear Determinism

Anonymous Author(s)
Affiliation
Address
email

Abstract

Probabilistic inference in many real-world problems often requires graphical models with (1) nonlinear deterministic constraints between random variables (e.g., laws of physics) and (2) piecewise distributions (e.g., priors enforcing bounded parameter values). While Gibbs sampling provides an attractive asymptotically unbiased MCMC approximation approach that does not require proposal design or tuning, it cannot be directly applied to models with determinism in the case of (1) and is difficult to automate in the case of (2). To address both limitations, we introduce a rich class of piecewise algebraic graphical models with nonlinear determinism, where we show that deterministic constraints can always be eliminated, thus permitting the application of Gibbs sampling; we further show the collapsed model always permits one symbolic integral — sufficient to automatically derive conditionals for Gibbs sampling. We evaluate this fully automated Symbolic Gibbs sampler on models motivated by physics and engineering and show it converges an order of magnitude faster than existing Monte Carlo samplers.

1 Introduction

Recently, representations of distributions by some forms of piecewise functions have attracted attention for their (1) closed-form analytical properties and (2) ability to approximate arbitrary distributions up to arbitrary precision. To date, such forms have been restricted to the piecewise polynomials/exponential with hyperrectangular, hyper-rhombus and linear partitioning constraints [SW11, She12, SA12].

In this paper we introduced a rich class of *polynomial piecewise fractional functions* (PPFs), that is, the mixture of fractional algebraic functions with polynomial partitioning constraints. To the best of our knowledge, this is the richest class of piecewise functions studied in the literature of graphical models and probabilistic inference so far.

As we will show, a large subset of this family (1) remain closed under operations such as algebraic fractional substitutions. (2) have closed-form univariate integrals. The former property makes PPFs an important tool in fully-automated reasoning conditioned on (non)linear deterministic constraints between model random variables (see Section2) and the latter property makes them an appropriate application of *exact* Gibbs sampling. In the sense that the univariate CDF computations required in each step of Gibbs sampling can be computed in closed-form and prior to the sampling process. In this way, we create a *symbolic Gibbs* algorithm that saves a significant amount of computation by avoiding per-sample computations, and shows dramatically improved performance compared to existing samplers on complex models motivated by physics and engineering. The combination of these novel contributions should make probabilistic reasoning applicable to variety of new appli-

cations that, to date, have remained beyond the tractability and accuracy boundaries of existing inference methods.

2 Observed Deterministic Constraints (Determinism)

Observed determinism, can appear in applications such as failure detection [HNV04] where random variables $X_1, \dots X_k$ are not directly observable but a function $f(X_1, \dots, X_k)$ of them may be observed (indirect measurement). Some families of distributions such as *conditional linear Gaussians* (CLGs) [LJ01] are closed under linear transformations and consequently allow linear determinism. In [LRR13] the restriction is imposed on the network topology rather than the densities. They can only handle the observed summation of variables on a very simple Bayesian network. The generalization of this technique does not seem straight-forward (if possible).

Handling nonlinear determinism is much trickier. Methods based on Hamiltonian MC are used for sampling under particular constraints [Har08] but they perform poorly in piecewise models (due to poor approximation of Hamiltonian dynamism in discontinuities)[Nea11]. [CS05] approximate non-linear determinism with piecewise linear constraints via dividing the space into hypercubes. This method can hardly be used in dimensions more than 2 or 3 since the number of partitions required to preserve approximation accuracy is exponential in dimensionality. The solution often suggested by probabilistic programming languages is to approximate the observed determinism via adding noise to the observation (hard to soft constraint conversion via measurement error). But in practice if the added noise is large, the approximation bounds may become arbitrarily large and if the noise is small, the mixing rate may become extremely slow (near-deterministic problem) [CC87]. In short, inference conditioned on nonlinear determinism is still an open problem [LRR13] and even for linear constraints, the existing works are too restrictive.

2.1 Collapsing Determinism

To express a deterministic constraint $f(x_1, \dots, x_n) = c$, we assume that in the variable set over which the probability measure is defined, there exist a random variable Z such that $p(Z = z|x_1, \dots, x_n) = \delta[f(x_1, \dots, x_n) - z]$.[1] ,[2] Therefore, the constraint corresponds the event $Z = c$.

In the following theorem, we use the calculus of Dirac delta and generalize the concept of change of random variables to (not necessarily) reversible functions $f(x_1, \cdot)$. Since in formula 1 one variable is collapsed i.e. marginalized out of from the conditional joint, we refer to it as *dimension reduction*.

Theorem 1 (Dimension reduction)**.** *Let $p(Z = z|x_1, \dots, x_n) = \delta[f(x_1, \dots, x_n) - z]$, where $f(x_1, \dots, x_n) = 0$ has real and simple roots for x_1 with a non-vanishing continuous derivative $\partial f(x_1, \dots, x_n)/\partial x_1$ at all those roots. Denote the set of all roots by $\mathcal{X}_1 = \{x_1 \mid f(x_1, \dots, x_n) - z = 0\}$. (Note that each element of $\mathcal{X}_1$ is a function of the remaining variables $x_2, \dots, x_n, z$.) Then:*

$$p(x_2, \dots, x_n \,|\, Z = z) \propto \sum_{x_1^i \in \mathcal{X}_1} \frac{p(X_1 = x_1^i, x_2, \dots, x_n)}{\left| \big(\partial f(x_1, \dots, x_n)/\partial x_1\big)|_{x_1 \leftarrow x_1^i} \right|} \tag{1}$$

Proof. $p(x_2, \dots, x_n \,|\, Z = z)$ is proportional to

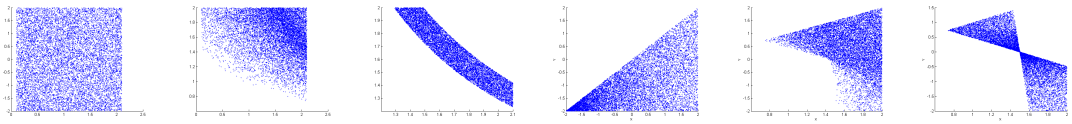
$$\int_{-\infty}^{\infty} p(x_1, \dots, x_n) p(Z = z \,|\, x_1, \dots, x_n) \, \mathrm{d}x_1 = \int_{-\infty}^{\infty} p(x_1, \dots, x_n) \delta\big(f(x_1, \dots, x_n) - z\big) \, \mathrm{d}x_1 \tag{2}$$

According to [GS64] there is a unique way to define the composition of Dirac delta with an arbitrary function $h(x)$:

$$\delta(h(x)) = \sum_i \frac{\delta(x - r_i)}{|\partial h(x)/\partial x|} \tag{3}$$

[1] This is to prevent Borel-Kolmogorov paradox [Kol50] that arises when conditioning is on an event of zero probability without specifying the random variable it is drawn from.

[2] $\delta\big(f(\cdot) - z\big)$ should be thought of as a limit of a normal distribution centered at $f(\cdot)$ and a variance that tends to zero.



(a) (b) (c) (d) (e) (f)

Figure 1: Prior/posterior joint distributions of pairs of random variables in the *collision* example. (a) $p(M_1, V_1)$, (b) $p(M_1, V_1 \mid P_{\text{tot}} = 3)$, (c) $p(M_1, V_1 \mid P_{\text{tot}} = 3, V_2 = 0.2)$, (d) $p(V_1, V_2)$, (e) $p(V_1, V_2 \mid P_{\text{tot}} = 3)$, (f) $p(V_1, V_2 \mid M_1 = 2, P_{\text{tot}} = 3)$ using rejection sampling on the δ-collapsed model.

where r_i are all (real and simple) roots of $h(x)$ and $h(x)$ is continuous and differentiable in the root points. By (2), (3) and *Tonelli's theorem*[3] $p(x_2, \ldots, x_n \mid Z = z) \propto$

$$\sum_{x_1^i \in \mathcal{X}_1} \frac{\int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)\delta(x_1 - x_1^i)\,\mathrm{d}x_1}{\left|\left(\partial f(x_1, \ldots, x_n)/\partial x_1\right)|_{x_1 \leftarrow x_1^i}\right|}$$

which implies (1). □

Reconstructing collapsed variables. If $f(x_1, \cdot)$ is reversible (w.r.t. x_1), given the realization (e.g. taken sample) of variables x_2, to x_n, the value of the *collapsed variable* x_1 is determined by $f^{-1}(z, \cdot)$. If not, sample value of x_1 is $x_1^i \in \mathcal{X}_\infty$ with a probability proportional to: $\big(p(x_1^i, x_2, \ldots, x_n)\delta(x_1 - x_1^i)\,\mathrm{d}x_1\big)/\left|\left(\partial f(x_1, \ldots, x_n)/\partial x_1\right)|_{x_1 \leftarrow x_1^i}\right|$. If more than one deterministic relationship exists, theorem 1 can be used multiple times and the collapsed variables are reconstructed in the reverse order they are eliminated.

To clarify the theorem and motivate the next section we provide an example:

Example (Collision model). Masses M_1 and M_2 with velocities V_1 and V_2 collide to form a single mass $(M_1 + M_2)$ with momentum $P_{\text{tot}} = M_1V_1 + M_2V_2$ (assuming that there is no dissipation). The masses and velocities are unknown but their prior distributions are given:[4]

$$p(M_1) = \mathcal{U}(0.1,\, 2.1), \quad p(M_2) = \mathcal{U}(0.1,\, 2.1), \quad p(V_1) = \mathcal{U}(-2,\, 2), \quad p(V_2 \mid V_1) = \mathcal{U}(-2,\, V_1)$$

Therefore, $p(M_1, M_2, V_1, V_2) = \begin{cases} \frac{1}{16V_1+32} & \text{if } 0.1<M_1<2.1, 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ 0 & \text{otherwise} \end{cases}$.

Suppose we observe $P_3 = 3$. To apply Theorem 1, we solve $(M_1V_1 + M_2V_2 - 3)$ w.r.t. a variable (say M_1 with the unique solution $(3 - M_2V_2)/V_1$). Since $\left|\frac{\partial(M_1V_1+M_2V_2)}{\partial M_1}\right| = |V_1|$, by (1): $p(M_2, V_1, V_2 \mid P_{\text{tot}} = 3)$ is proportional to $p(M_2, V_1, V_2) =$

$$\begin{cases} \frac{1}{V_1(16V_1+32)} & \text{if } 0<V_1, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ \frac{-1}{V_1(16V_1+32)} & \text{if } V_1<0, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \frac{1}{V_1(16V_1+32)} & \text{if } 0<V_1, 0.1V_1<3-M_2V_2<2.1V_1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ \frac{-1}{V_1(16V_1+32)} & \text{if } V_1<0, 2.1V_1<3-M_2V_2<0.1V_1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

By sampling from the polynomial piecewise fractional (PPF) function (4), different inference tasks can be carried out. Several such tasks are depicted in Figure 1. ◇

Figure 1 illustrates that even in this simple and low-dimension example, variable transformation can lead to complicated PPF patterns that do not resemble the bell-shaped densities often studied in the literature. Reasoning on the family of PPFs does provide a automated solution for inference in presence of a large set of (non)linear deterministic relationships between random variables. Despite the importance of the latter problem, it is clearly only an application of the former. Nonetheless, the family of PPFs and inference on them is not studied in the literature so far. In the rest of the paper, we will fill this gap.

[3] Tonelli's theorem says that for non-negative functions, sum and integral are interchangeable.

[4] $\mathcal{U}(a,\, b)$ denotes a uniform distribution on interval $[a, b]$.

3 Polynomial Piecewise Fractionals (PPFs)

A PPF is a function in form $f = \sum_{i=1}^{m} \mathbb{I}[\phi_i] \cdot f_i$ where $\mathbb{I}[\cdot]$ denotes the indicator function. Using expanded notation,

$$f = \begin{cases} f_1 & \text{if } \phi_1 \\ \vdots & \\ f_m & \text{if } \phi_m \end{cases} = \begin{cases} \frac{N_1}{D_1} & \text{if } \varphi_{1,1} \lesseqgtr 0,\ \varphi_{1,2} \lesseqgtr 0,\ \dots \\ \vdots & \\ \frac{N_m}{D_m} & \text{if } \varphi_{m,1} \lesseqgtr 0,\ \varphi_{m,2} \lesseqgtr 0,\ \dots \end{cases} \tag{5}$$

where each *sub-function* $f_i := \frac{N_i}{D_i}$ is a (multivariate) polynomial fraction and *conditions* ϕ_i partition the space of function variables. Each ϕ_i is a conjunctions of some inequalities ($\lesseqgtr$ stands for $>$ or $<$)[5] where each *atomic constraint* $\varphi_{i,j}$ is a polynomial. (6) shows, PPFs are closed under elementary operations.

$$\begin{cases} f_1 & \text{if } \phi_1 \\ f_2 & \text{if } \phi_2 \end{cases} \times \begin{cases} g_1 & \text{if } \psi_1 \\ g_2 & \text{if } \psi_n \end{cases} = \begin{cases} f_1 \times g_1 & \text{if } \phi_1, \psi_1 \\ f_1 \times g_2 & \text{if } \phi_1, \psi_2 \\ f_2 \times g_1 & \text{if } \phi_2, \psi_1 \\ f_2 \times g_2 & \text{if } \phi_2, \psi_2 \end{cases} \tag{6}$$

$$f|_{x \leftarrow \frac{F}{G}} = \begin{cases} f_1|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_1|_{x \leftarrow \frac{F}{G}} \\ \vdots & \\ f_m|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_m|_{x \leftarrow \frac{F}{G}} \end{cases} \tag{7}$$

They are also closed under polynomial fractional substitution (7). The reason is that in the r.h.s of (7), sub-functions $f_i|_{x \leftarrow \frac{F}{G}}$ are polynomial fractions (PFs) (which are closed under substitution). Conditions $\phi_i|_{x \leftarrow \frac{F}{G}}$ are fractional but as (8) shows, they can be restated as (multiple) case-statements with polynomial conditions.

$$\left(\begin{cases} f_1 & \text{if } \frac{H_1}{H_2} > 0 \\ \vdots & \end{cases} \right) = \begin{cases} f_1 & \text{if } H_1 > 0, H_2 > 0 \\ f_1 & \text{if } H_1 < 0, H_2 < 0 \\ \dots & \end{cases} \tag{8}$$

$$\left| \left(\begin{cases} \frac{N_1}{D_1} & \text{if } \phi_1 \\ \vdots & \end{cases} \right) \right| = \begin{cases} \frac{N_1}{D_1} & \text{if } N_1 > 0, D_1 > 0, \phi_1 \\ \frac{N_1}{D_1} & \text{if } N_1 < 0, D_1 < 0, \phi_1 \\ \frac{-N_1}{D_1} & \text{if } N_1 > 0, D_1 < 0, \phi_1 \\ \frac{-N_1}{D_1} & \text{if } N_1 < 0, D_1 > 0, \phi_1 \\ \dots & \end{cases} \tag{9}$$

Finally, their closure under *absolute value*, as seen in (9), means PPFs are closed under all operations utilized in Theorem 1.

3.1 Analytic integration

In general, PPFs are not closed under integration; however, a large subset of them have closed-form single variable integrals. We focus on the following fairly expressive subset of PPFs:

PPF*. A PPF* is a PPF in which:

1. Each atomic constraint $\varphi_{i,j}$ can be written as a product of some terms $t_{i,j,k}$ where the maximum degree of each variable in each $t_{i,j,k}$ is less or equal to 2.
2. The sub-function denominators, D_i, can be factorized into (not necessarily distinct) polynomials $D_{i,l}$ where the maximum degree of each variable in each $D_{i,l}$ is less or equal to 2.

Here is an example of a PPF* case-statement:

$$\frac{x^2y^3 + 7xz + 10}{(5xy^2 + 2)(y + x)^3} \quad \text{if } (y^2 + z^2 - 1)(x^2 + 2xy) > 0 \tag{10}$$

Analytic univariate PPF* integration. Now we provide a procedure to perform integration on PPF* functions.

It can be shown that if in a PPF* all variables except one are instantiated, the resulting univariate function has a closed form integral. To perform exact Gibbs sampling, this is sufficient because in each step of Gibbs sampling only one variable is uninstantiated.

[5] We do not define the value of piecewise density functions on their partitioning hyperplanes and do not allow $\delta(\cdot)$ potentials have roots on the partitions.

However, we want to go a step further and compute univariate integrals of multivariate piecewise functions *without* instantiating the remaining variables.

This may look impossible since in the latter case, the integration bounds depend on the values of uninstantiated conditions. But as the following procedure shows, it is indeed possible for the PPF* class:

Suppose $\int_\alpha^\beta f \, \mathrm{d}x$ is intended where f is a PPF*.

1. *(Partitioning).* The integral of the piecewise function f is the summation of its case statement integrals:
$$\int \sum_{i=1}^{m} \mathbb{I}[\phi_i] \cdot f_i \, \mathrm{d}x = \sum_{i=1}^{m} \int \mathbb{I}[\phi_i] \cdot f_i \, \mathrm{d}x$$
Therefore we only need to show that a single PPF* case-statement is integrable.

2. *(Canonicalization).* A PPF* case statement can be restated in the form of multiple case statements in which the degree of each variable in each atomic constraint is at most 2. For instance, (10) can be restated as:
$$\begin{cases} \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1)>0, (x^2+2xy)>0 \\ \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1)<0, (x^2+2xy)<0 \end{cases} \tag{11}$$

3. *(Condition solution).* For the integration variable x, a PPF* case statement can be transformed into a piecewise structure with atomic constraints in form $x > L_i$ or $x < U_i$ or $I_i > 0$, where L_i, U_i and I_i are algebraic expressions (not necessarily polynomials) that do not involve x.
For instance, if expressions A, B and C do not involve x, the case statement (12) is replaced by cases statements (13).
$$f_1 \quad \text{if } (A \cdot x^2 + B \cdot x + C) > 0 \tag{12}$$
$$\begin{cases} f_1 & \text{if } (A>0),\ (x > \frac{-B+\sqrt{B^2-4AC}}{2A}) \\ f_1 & \text{if } (A>0),\ (x < \frac{-B-\sqrt{B^2-4AC}}{2A}) \\ f_1 & \text{if } (A<0),\ (x > \frac{-B-\sqrt{B^2-4AC}}{2A}), (x < \frac{-B+\sqrt{B^2-4AC}}{2A}) \end{cases} \tag{13}$$
3. (Bounding). The bounded integral of a case statement associated with $\{L_i\}_i$, $\{U_i\}_i$ and $\{I_i\}_i$ is itself a case-statement with the same independent constraints, lower bound LB $=\max\{\alpha, L_i\}$ and upper bound UB $=\min\{\beta, U_i\}$. For example:
$$\int_\alpha^\beta \Big[x^3+xy \quad \text{if } (x>\mathbf{3}), (x>\mathbf{y+1}), (x<\mathbf{y^2-7}), (\mathbf{-y/z>1}), (\mathbf{y>0})\Big] \, \mathrm{d}x =$$
$$\Big[\int_{\max\{\alpha,\,\mathbf{3},\,\mathbf{y+1}\}}^{\min\{\beta,\,\mathbf{y^2-7}\}} x^3+xy \, \mathrm{d}x\Big] \quad \text{if } (-\frac{\mathbf{y}}{\mathbf{z}} > \mathbf{1}) \wedge (\mathbf{y>0})$$
4. (sub-function integration). What is remained is to compute infinite integral of sub-functions. The restrictions imposed on PPF* sub-functions guarantee that they have closed-form univariate integrals. These integrals are computed by performing polynomial division (in case the degree of x in the sub-function's numerator is more than its denominator), followed by partial fraction decomposition and finally, using a short list of indefinite integration rules.

4 Symbolic Gibbs Sampling

Our *symbolic Gibbs sampling* is based on a simple but significantly useful insight: If $p(X_1, \ldots, X_N)$ has analytic integrals w.r.t. all variables X_i (as is the case with PPF* densities), then the costly CDF computations can be done *prior to the sampling process rather than per sample*. It is sufficient to construct a mapping $\mathcal{F}$ from variables X_i to their corresponding (unnormalized conditional) analytical CDFs.
$$\mathcal{F}\colon \{X_1, \ldots X_N\} \to (\mathbb{R}^N \to \mathbb{R}^+ \cup \{0\})$$
$$X_i \mapsto \int_{-\infty}^{X_i} p(X_i = t, \mathbf{X}_{-i}) \, \mathrm{d}t \tag{14}$$

Table 1: Parameters corresponding each experimental model

#	Experiment	HMC	SMC	Evidence
1	collision	$\sigma^2_{P_t} = 0.05$	$\sigma^2_{P_t} = 0.1$	$P_t = 1.5n$
2	power line	$\sigma^2_G = 0.02$	$\sigma^2_G = 0.07$	$G = n/10.17$

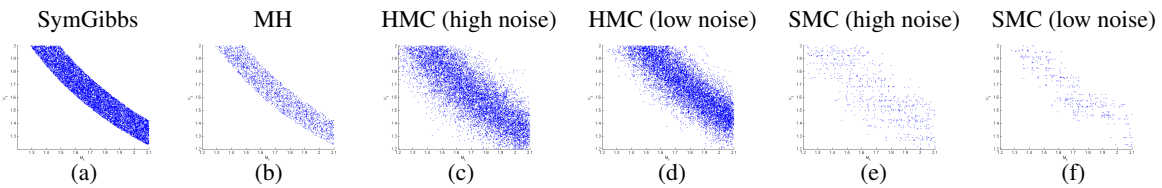


Figure 2: 10000 samples taken from the distribution Fig. (1-c) using (a) Symbolic Gibbs sampler and(b) MH with *proposal variance* 0.8 on the reduced-dimension model as well as (c) Hamiltonian Monte Carlo (HMC) with a measurement error variance 0.2, , (d) and 0.01 as well as Anglican implementation of SMC alg. with parameters (e) $\sigma^2_{V_2} = 0.01$, $\sigma^2_{P_{\text{tot}}} = 0.2$ and (f) $\sigma^2_{V_2} = 0.01$, $\sigma^2_{P_{\text{tot}}} = 0.1$ on the *approximated-by-noise* model.

Note that the difference between (18) and (14) is that in the former, all variables except X_i are already instantiated therefore $\text{CDF}(X_i \,|\, \mathbf{x}_{-i})$ is a univariate function but $\mathcal{F}$ is N-variate since variables $\mathbf{X}_{-i}$ are kept uninstantiated and symbolic. Provided with such a map, in the actual sampling process, to sample $x_i \sim p(X_i \,|\, \mathbf{x}_{-i})$, it is sufficient to instantiate the analytical CDF associated to X_i with $\mathbf{x}_{-i}$ to obtain the appropriate univariate conditional CDF (see Algorithm 1). This reduces the number of CDF computations from $N \cdot T$ to N where T is the number of taken samples.

If CDF inversion (required for inverse transform sampling) is also computed analytically, then Gibbs sampling may be done fully analytically. However, analytical inversion of PPF*s can be very complicated and instead in the current implementation, we approximate the CDF^{-1} computation via *binary search*. This requires several function evaluations per sample. Nonetheless, (unlike integration), function evaluation is a very fast operation. Therefore, this suffices for highly efficient Gibbs sampling as will show experimentally in the next section.

Algorithm 1 SYMBOLICGIBBS$(\mathbf{X}, \mathbf{x}^{(0)}, \mathcal{F}, T)$

Input: random variables $\mathbf{X} := \langle X_1, \ldots, X_N \rangle$; initialization $\mathbf{x}^{(0)} := \langle x_1^{(0)}, \ldots, x_N^{(0)} \rangle$; a mapping $\mathcal{F}$ from $\mathbf{X}$ to analytical conditional CDFs; desired number of output samples T.
Output: a sequence of samples.
for $t = 1, 2, \ldots T$ **do**
  $\langle x_1^{(t)}, \ldots, x_N^{(t)} \rangle \leftarrow \langle x_1^{(t-1)}, \ldots, x_N^{(t-1)} \rangle$
  for each $X_i \in \mathbf{X}$ **do**
    $F(\mathbf{X}) \leftarrow \mathcal{F}(X_i)$ *//analytic conditional CDF w.r.t. X_i*
    $\text{CDF}(X_i) \leftarrow F(x_1^{(t)}, \ldots, x_{i-1}^{(t)}, X_i, x_{i+1}^{(t)}, \ldots, x_N^{(t)})$
    //Inverse transform sampling:
    $u \sim \mathcal{U}(0, \text{Cdf}(\infty))$
    $x_i^{(t)} \leftarrow \text{CDF}^{-1}(u)$
  end for
end for
Return $\left\langle \langle x_1^{(1)}, \ldots, x_N^{(1)} \rangle, \ldots, \langle x_1^{(T)}, \ldots, x_N^{(T)} \rangle \right\rangle$

5 Experimental Results

We (a) compare the performance of the proposed symbolic Gibbs against other MCMC methods on models with observed determinism. Moreover, we are interested in (b) studying the performance of sampling if instead of collapsing determinism, constraints are relaxed by noise (as often done in probabilistic programming toolkits).

Compared algorithms. We compare the proposed *symbolic Gibbs sampler* (SymGibbs) to *baseline Gibbs* (BaseGibbs) [Pea87], *rejection sampling* (Rej) [HH64], *tuned Metropolis-Hastings* (MH) [RGG+97], *Hamiltonian Monte Carlo* (HMC) using Stan probabilistic programming language

[Sta14] and *Sequential Monte Carlo* (SMC) using Anglican probabilistic programming language [WvdMM14].[6] SymGibbs and BaseGibbs require no tunings. MH is automatically tuned after [RGG$^+$97] by testing 200 equidistant proposal variances in interval $(0, 0.1]$ and accepting a variance for which the acceptance rate closer to 0.24.

SymGibbs, BaseGibbs and MH are run on *collapsed-determinism* models while in the case of HMC and SMC, determinism is softened by observation noise. It should be mentioned that the state-of-the-art probabilistic programming languages, disallow deterministic relationships among continuous random variables be observed.[7] The solution that these off-the-shelf inference frameworks often suggest (or impose) is to approximate the observed determinism via adding noise to the observation [PHF10].[8] To soften the determinism in HMC and SMC, the observation of a deterministic variable Z is approximated by observation of a newly introduced variable with a Gaussian prior centered at Z and with noise variance (parameter) σ_Z^2. Anglican's syntax requires adding noise to all observed variables. Therefore, in the case of SMC, stochastic observations are also associated with noise parameters. All used parameters are summarized in Table 1. SymGibbs, BaseGibbs, Rej and MH have single thread java implementations. The number of threads and other unspecified parameters of Stan and Anglican are their default settings. All algorithms run on a 4 core, 3.40GHz PC.

Measurements. To have an intuitive sense of the performance of the different MCMCs, Figure 2 depicts 10000 samples are taken from the posterior of Figure 1-c using the introduced sampling algorithms.

For quantitative comparison, in each experiment, all non-observed stochastic random variables of the model form the query vector $\mathbf{Q} = [Q_1, \ldots, Q_\zeta]$. The number of samples taken by a Markov chain Γ up to a time t is denoted by n_Γ^t and the samples are denoted by $\mathbf{q}_\Gamma^{(1)}, \ldots, \mathbf{q}_\Gamma^{(n_\Gamma^t)}$ where $\mathbf{q}_\Gamma^{(i)} := [q_{1,\Gamma}^{(i)}, \ldots, q_{\zeta,\Gamma}^{(i)}]$

We measure mean absolute error (MAE) (15) vs (wall-clock) time t where $\mathbf{q}^* := [q_1^*, \ldots q_\zeta^*]$ is the ground truth mean query vector (that is computed manually due to the symmetry of the chosen models).

$$\mathrm{MAE}_\Gamma(t) := \frac{1}{\zeta \cdot n_\Gamma^t} \sum_{j=1}^{\zeta} \sum_{i=1}^{n_\Gamma^t} \left| q_{j,\Gamma}^{(i)} - q_j^* \right| \tag{15}$$

In each experiment and for each algorithm, $\gamma = 15$ Markov chains are run, and for each time point t, average and standard error of $\mathrm{MAE}_1(t)$ to $\mathrm{MAE}_\gamma(t)$ are plotted.

5.1 Experimental models

Multi-object collision model. Consider a variation of the collision model in which n objects collide. Let all V_i and M_i share a same uniform prior $U(0.2, 2.2)$ and the constraint be $\sum_{i=1}^n M_i V_i = P_{\text{tot}}$. The symmetry enables us to compute the posterior ground truth means values manually:

$$M^* = V^* = \sqrt{P_{\text{tot}}/n} \tag{16}$$

Conditioned on $P_{\text{tot}} = 1.5n$, all masses M_i and velocities V_i are queried. By (16), all elements of the ground truth vector $\mathbf{q}^*$ are $\sqrt{1.5}$. MAE vs. time is depicted in Figures 3.a & b for a 10-D and a 30-D model, respectively.

Building wiring model. An electrical circuit composed of n, $10\Omega \pm 5\%$ parallel resistor elements R_i (with priors $p(R_i) = U(9.5,\ 10.5)$). The resistors are inaccessible i.e. the voltage drop and

[6] We also tested the other algorithm implemented by Anglican, namely *Particle-Gibbs* (PGibbs) (a variation of Particle-MCMC[ADH10]) and *random database* (RDB) (an MH-based algorithm introduced in [WSG11]) (see [WvdMM14]). In our experimental models, the performance of these algorithms is very similar to (SMC). Therefore, for the readability of the plots, we did not depict them.

[7] In BUGS [LSTB09], *logical nodes* cannot be given data or initial values . In PyMC [PHF10] deterministic variables have no *observed flag*. In Stan [Sta14] if you try to assign an observation value to a deterministic variable, you will encounter an error message: "attempt to assign variable in wrong block" while Anglican [WvdMM14] throws error "invalid-observe", etc.

[8] For example in the collision model, the observation $P_{\text{tot}} = 3$ would be approximated with a normal distribution $\mathcal{N}(P_{\text{tot}} - 3, \sigma_\eta^2)$ where the variance σ_η^2 is the noise parameter.

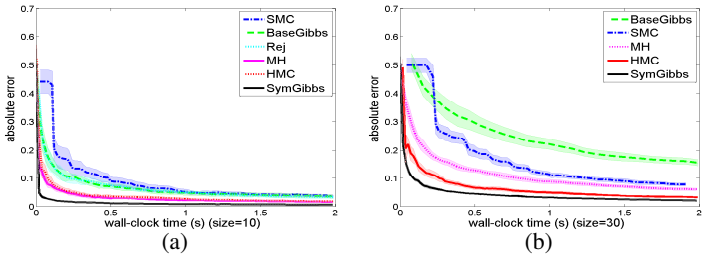


Figure 3: MCMC Convergence measurements in the symmetric multi-object collision model: Absolute error vs time for collision of (a) 4 and (b) 20 objects.

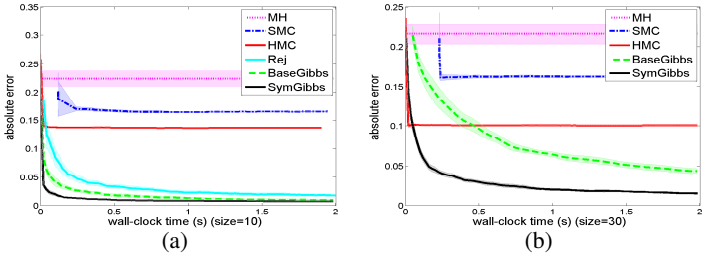


Figure 4: MCMC Convergence measurements in the building wiring model: Absolute error vs time for a model with (a) size 4 (i.e. 4 paralleled resistors) and (b) size 30.

the current associated with them cannot be measured directly. Given the source voltage V and the total input current I, the posterior distribution of the element resistances are required. Here the deterministic constraint is:

$$\frac{1}{R_1} + \ldots + \frac{1}{R_n} = c \tag{17}$$

where $c = \frac{I}{V}$. Equations if the form (17) are generally referred to as *reduced mass* relationships and have applications in the electrical, thermal, hydraulic and mechanical engineering domains.

Let the observation be $c = 3n/(2 * 10.5 + 9.5)$. Due to the symmetry of the problem, the posterior ground truth mean is known:

$$R_i^* = \frac{n}{c} = 10.166667 \qquad \text{for } i = 1, \ldots, n$$

MAE vs. time for networks of 10 and 30 resistors are depicted in Figures 4.a & b respectively.

Experimental evaluations. Plots of Figure 2 shows that MH and SMC suffer from low *effective sample size*. Note that the apparent sparsity of plots 2-b, 2-e & 2-f is due to repeated samples (rejected proposals). The carried out quantitative measurements (Figures 3 and 4) indicate that in all experimental settings, symbolic Gibbs constantly and significantly performs the best.

All quantitative measurements (Figures 3 and 4) indicate that hard to soft constraint conversion (via introducing measurement error) ends in poor results. Interestingly, in the Building wiring model, even in a dimensionality as low as 10, the Metropolis-Hasting based algorithms (i.e. HM, HMC and SMC) may not converge to the (manually computed) ground truth or their convergence rate is extremely low. This happens regardless of the way determinism is handled.

6 Conclusions

In this paper we introduced a rich class of polynomial piecewise fractional functions (PPFs) using polynomial partitioning constraints. To the best of our knowledge, this is the richest class of piecewise functions studied in the literature of graphical models and probabilistic inference so far. We showed that this family is expressive enough to remain closed under operations required to transform networks with nonlinear deterministic constraints to purely stochastic PPF models amenable to Gibbs sampling.

We showed that a large subset of PPFs have symbolic univariate integrals, which together with determinism elimination, enabled the main contribution of this paper: a fully-automated exact Gibbs sampler called *symbolic Gibbs*. In *symbolic Gibbs*, all univariate CDFs required for Gibbs sampling are computed analytically and prior to the actual sampling process. In this way, *symbolic Gibbs* saves a significant amount of computation by avoiding per-sample computations, and shows dramatically improved performance compared to existing samplers on complex models motivated by physics and engineering. The combination of these novel contributions should make probabilistic reasoning applicable to variety of new applications that, to date, have remained beyond the tractability and accuracy boundaries of existing inference methods.

References

[ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[CC87] Homer L Chin and Gregory F Cooper. Bayesian belief network inference using simulation. In *UAI*, pages 129–148, 1987.

[CS05] Barry R Cobb and Prakash P Shenoy. Nonlinear deterministic relationships in Bayesian networks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 27–38. Springer, 2005.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[GS64] IM Gel'fand and GE Shilov. Generalized functions. vol. 1: Properties and operations, fizmatgiz, moscow, 1958. *English transl., Academic Press, New York*, 1964.

[GW92] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[Har08] Carsten Hartmann. An ergodic sampling scheme for constrained hamiltonian systems with applications to molecular dynamics. *Journal of Statistical Physics*, 130(4):687–711, 2008.

[HH64] John Michael Hammersley and David Christopher Handscomb. *Monte Carlo methods*, volume 1. Methuen London, 1964.

[HNV04] Arne B Huseby, Morten Naustdal, and Ingeborg D Varli. System reliability evaluation using conditional Monte Carlo methods. *Preprint series. Statistical Research Report*, 2, 2004.

[Kol50] Andrei Nikolaevich Kolmogorov. Foundations of the theory of probability. 1950.

[LJ01] Steffen L Lauritzen and Frank Jensen. Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11(2):191–203, 2001.

[LRR13] Lei Li, Bharath Ramsundar, and Stuart Russell. Dynamic scaled sampling for deterministic constraints. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 397–405, 2013.

[LSTB09] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

[Nea11] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

[Pea87] Judea Pearl. Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence*, 32(2):245–257, 1987.

[PHF10] Anand Patil, David Huard, and Christopher J Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software*, 35(4):1, 2010.

[RGG+97] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[SA12] Scott Sanner and Ehsan Abbasnejad. Symbolic variable elimination for discrete and continuous graphical models. In *AAAI*, 2012.

[She12] Prakash P Shenoy. Two issues in using mixtures of polynomials for inference in hybrid Bayesian networks. *International Journal of Approximate Reasoning*, 53(5):847–866, 2012.

[Sta14] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0*, 2014.

[SW11] Prakash P Shenoy and James C West. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52(5):641–657, 2011.

[WSG11] David Wingate, Andreas Stuhlmueller, and Noah D Goodman. Lightweight implementations of probabilistic programming languages via transformational compilation. In *International Conference on Artificial Intelligence and Statistics*, pages 770–778, 2011.

[WvdMM14] Frank Wood, Jan Willem van de Meent, and Vikash Mansinghka. A new approach to probabilistic programming inference. In *Proceedings of the 17th International conference on Artificial Intelligence and Statistics*, 2014.

Gibbs sampling: In Gibbs sampling [GG84] drawing a sample for $\mathbf{X} = \{X_1, \ldots, X_N\}$ takes place in N steps. In the i-th step, X_i is sampled conditioned on the last realization of the others: $x_i \sim p(X_i \,|\, \mathbf{x}_{-i})$. To perform this task, the following univariate (conditional) *cumulative distribution function* (CDF) is computed by (18) and samples are taken via inverse transform sampling.

$$\mathrm{CDF}(X_i \,|\, \mathbf{x}_{-i}) \propto \int_{-\infty}^{X_i} p(X_i = t,\, \mathbf{X}_{-i} = \mathbf{x}_{-i})\, \mathrm{d}t \tag{18}$$

Gibbs sampling does not require any tuning and since it directly samples from the distribution (rather than indirectly and through proposals) it has high effective sample size. However, in general, closed-form computation of the CDF integrals is not possible and approximations (such as [GW92]) are costly. Considering that N univariate integrals should be computed per sample, such approximate Gibbs samplers are typically slow.