Closed-form Gibbs Sampling for Graphical Models with (Non)Linear Constraints

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 often requires manual derivation of complex piecewise conditional distributions 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 through collapsing. We further show that the form of the collapsed model always permits one *symbolic* integral — sufficient to *automatically* derive conditionals for Gibbs sampling. We evaluate this fully automated Symbolic Gibbs sampler for nonlinear piecewise graphical models on examples motivated by physics and engineering and show it converges an order of magnitude faster than existing Monte Carlo samplers.

Introduction

Probabilistic inference in many real-world problems often requires graphical models with (non)linear constraints between random variables. Such constraints appear in case instead of direct observation of random variables, deterministic functions of them are observed. To illustrate this, consider the following running example (with the graphical model of Figure 1):

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 total momentum $P_{tot} = M_1V_1 + M_2V_2$ (assuming that there is no dissipation). The prior density of masses and velocities are:*[1]

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

The total momentum is observed to be 3.0.

$$P_{tot} = M_1V_1 + M_2V_2 = 3.0 \tag{1}$$

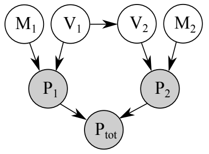


Figure 1: Bayesian network of the *collision model*. Shaded circles correspond random variables that are functions of other variables.

In such problems, the support of posterior densities are (non)linear sub-manifolds of the parameter space (e.g. hyperplane (1) in the collision model). Carrying out inference on such models is by no means trivial (Pen06).

To evade the complications, state-of-the-art MCMC based probabilistic inference tools suggest adding noise to the observation (hard to soft constraint conversion) (WvdMM14). Nonetheless, this strategy does not help much: If the added noise is large then the approximation bounds can be arbitrarily large and if it is small, the sampling mixing rate can be arbitrarily slow (LRR13; CC87).

The other potential solution is to reduce the dimensionality of the posterior distribution via Jacobian-based random variable transformations. Measure theoretic subtleties aside, such transformations are only applicable when the observed function is invertible with respect to at least one random variable. Using the properties of the Dirac delta, our first contribution is to propose a dimension reduction method that is more general in the sense that the observed function is not required to be invertible but should be solvable with one or several distinct roots. Up to our knowledge, this is the first time that Dirac delta is used in this context. Nonetheless, dimension reduction (either carried out via Jacobian-based or Dirac delta-based mechanism) does not completely eliminate the problem since as it will be shown shortly, the produced low dimensional distributions are highly piecewise and multimodal.

Exact inference on such models is almost never possible and the convergence rate of the approximate alternatives can be extremely low. For instance, the leapfrog mechanism by which Hamiltonian Monte Carlo (HMC) simulates the Hamiltonian dynamics relies on the assumption of smoothness (Nea11). This assumption does not hold in the adjacency of discontinuities (borders of pieces) leading to low proposal acceptance rates and poor performance. Slice sampling suffers from the multimodal nature of the distributions

[1] $\mathcal{U}(a,\, b)$ denotes a uniform density with support $[a, b]$.

that are in the focus of the present work. Similarly, near the borders of partitions, the acceptance rate of Metropolis-Hastings(MH) is typically low since in such areas the difference (e.g. KL-divergence) between MHs *proposal distribution* and the suddenly varying target distribution is often significant. The exception is Gibbs sampling. The latter method can be regarded as a particular variation of MH where the proposals are directly chosen from the target distribution and therefore follow the target distribution changes and multimodalities. Nonetheless, Gibbs samplers can be quite slow since the per sample computation conditionals that Gibbs relies on are costly and in general cannot be performed in closed form.

As such, in the second part of the paper we address the problem of sampling from highly piecewise distributions. We firstly introduce a reach class of *piecewise fractional functions* as a building block for piecewise graphical models.We show that this class is closed under the operations required for Dirac delta-based dimension reduction. This class is rich enough to approximate arbitrary density functions up to arbitrary precision. We further show that the form of resulting collapsed model always permits one closed-form integral – sufficient to analytically derive conditionals for Gibbs sampling prior to the sampling process which saves a tremendous amount of computations. We evaluate this fully-automated sampler for models motivated by physics and engineering and show it converges at least an order of magnitude faster than existing MCMC samplers, thus enabling probabilistic reasoning in a variety of applications that, to date, have remained beyond the tractability and accuracy purview of existing inference methods.

Preliminaries

Graphical models. Let $\mathbf{X} = \{X_1, \dots, X_N\}$ be a set of random variables with realizations in the form $\mathbf{x} = \{x_1, \dots, x_N\}$.[2] For the sake of notational consistency, throughout we assume $\mathbf{X}$ only contain continuous variables. To cover both directed and undirected graphical models we use *factor graph* notation (KFL01) and represent a joint probability density $p(\mathbf{X})$ in a factorized form as follows:

$$p(\mathbf{X}) \propto \prod_{\Psi_k \in \mathbf{\Psi}} \Psi_k(\mathbf{X}_k) \tag{2}$$

where $\Psi(\cdot)$ are non-negative *potential functions* of subsets $\mathbf{X}_k$ of $\mathbf{X}$.

Inference. The inference task studied in this paper is to compute the *posterior* joint density $p(\mathbf{Q} \,|\, \mathbf{E} = \mathbf{e})$ of a subset $\mathbf{Q}$ (*query*) of $\mathbf{X}$ conditioned on (realization $\mathbf{e}$ of) variables $\mathbf{E} \subset \mathbf{X} \backslash \mathbf{Q}$ (*evidence*):

$$p(\mathbf{Q} \,|\, \mathbf{E} = \mathbf{e}) \propto \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\mathbf{Q}, \mathbf{W} = \mathbf{w}, \mathbf{E} = \mathbf{e}) \, \mathrm{d}\mathbf{w} \tag{3}$$

where $\mathbf{W} = \{W_1, \dots, W_m\} := \mathbf{X} \backslash (\mathbf{Q} \cup \mathbf{E})$.

The integrals required in (3) are often intractable and hence we must often resort to MCMC methods such as Gibbs sampling (GG84) — the focus of this work.

Gibbs sampling. In this method drawing a sample for $\mathbf{X}$ 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 density function* (CDF) is computed by (4) and samples are taken via inverse transform sampling.

$$\text{CDF}(X_i \,|\, \mathbf{x}_{-i}) \propto \int_{-\infty}^{X_i} p(X_i = t, \mathbf{X}_{-i} = \mathbf{x}_{-i}) \, \mathrm{d}t \tag{4}$$

Observed Constraints

To express an observed constraint $f(x_1, \dots, x_n) = c$, we assume that in the variable set over which the probability measure is defined, there exists a random variable Z such that $p(Z = z | x_1, \dots, x_n) = \delta[f(x_1, \dots, x_n) - z]$.[3]

In the following theorem, we use the calculus of Dirac deltas and generalize the concept of change of random variables to (not necessarily) reversible functions $f(x_1, \cdot)$. Since in formula (5) one variable is collapsed i.e. marginalized out we refer to it as *dimension reduction*.

Theorem 1 (Dimension reduction)**.** *Let,*

$$p(Z = z | x_1, \dots, x_n) = \delta\big(f(x_1, \dots, x_n) - z\big)$$

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:*

$$p(x_2, \dots, x_n \,|\, Z = z) \propto \sum_{x_1^i \in \mathcal{X}_1} \frac{p(X_1 = x_1^i, x_2, \dots, x_n)}{\Big|\big(\partial f(x_1, \dots, x_n)/\partial x_1\big)|_{x_1 \leftarrow x_1^i}\Big|} \tag{5}$$

Proof. $p(x_2, \dots, x_n \,|\, Z = z) \propto$

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

$$= \int_{-\infty}^{\infty} p(x_1, \dots, x_n) \delta\big(f(x_1, \dots, x_n) - z\big) \, \mathrm{d}x_1 \tag{6}$$

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

$$\delta(h(x)) = \sum_i \frac{\delta(x - r_i)}{|\partial h(x)/\partial x|} \tag{7}$$

where r_i are all (real and simple) roots of $h(x)$ and $h(x)$ is continuous and differentiable in the root points. By (6), (7) and *Tonelli's theorem*[4] $p(x_2, \dots, x_n \,|\, Z = z) \propto$

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

which implies (5). □

[2] In case there is no ambiguity, we do not distinguish between random variables and their realizations; e.g., we abbreviate $p(X_i = x_i)$ by $p(x_i)$.

[3] This is to prevent Borel-Kolmogorov paradox (Kol50) that arises when conditioning is on an event with a probability that tends to zero without specifying the random variable it is drawn from. $\delta\big(f(\cdot) - z\big)$ should be thought of as a limit of a normal distribution centered at $f(\cdot)$ and a variance that tends to zero.

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

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

....

Therefore, $p(M_1, M_2, V_1, V_2)$ is equal to

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

....

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 (5), $p(M_2, V_1, V_2 \mid P_{\text{tot}} = 3)$ is proportional to

$$\begin{cases} \frac{1}{V_1(16V_1+32)} & \text{if } 0<V_1, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ \frac{-1}{V_1(16V_1+32)} & \text{if } V_1<0, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Using the joint distribution defined in the piecewise polynomial fractional (PPF) function of (8), we can evaluate various queries as depicted in Figure 2. ◇

Figure 2 illustrates that even in this simple and low-dimensional example, collapsing of nonlinear determinism can lead to complicated (posterior) distributions that do not resemble the smooth densities often studied in the literature. However, the use of PPFs as a building block for piecewise graphical models with nonlinear determinism does ensure that a large set of (non)linear deterministic relationships between random variables yield δ-collapsed models with factors still in the PPF family. The only task that remains then is to provide an automated sampling method for piecewise graphical models composed of an expressive class of PPFs, which we do in the next section.

Polynomial Piecewise Fractionals (PPFs)

Now that we have collapsed out determinism, we can apply Gibbs sampling. However, automated derivation of Gibbs sampling requires a univariate symbolic integral (required in (4)) which is not possible in the general case. We address this issue by introducing a rich family of distributions (namely PPFs for *polynomial piecewise fractional* functions) that remain closed under δ-collapsing and one symbolic integration — sufficient to automatically derive all conditional distributions required by Gibbs sampling in expressive piecewise graphical modelswith nonlinear determinism.

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

$$f = \begin{cases} f_1 & \text{if } \phi_1 \\ \vdots & \\ f_m & \text{if } \phi_m \end{cases} = \begin{cases} \frac{N_1}{D_1} & \text{if } \varphi_{1,1} \lesseqgtr 0,\ \varphi_{1,2} \lesseqgtr 0,\ \ldots \\ \vdots & \\ \frac{N_m}{D_m} & \text{if } \varphi_{m,1} \lesseqgtr 0,\ \varphi_{m,2} \lesseqgtr 0,\ \ldots \end{cases} \tag{9}$$

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 conjunction of some inequalities ($\lesseqgtr$ stands for $>$ or $<$)[5] where each *atomic constraint* $\varphi_{i,j}$ is a polynomial. (10) shows, PPFs are closed under elementary operations.

$$\begin{cases} f_1 & \text{if } \phi_1 \\ f_2 & \text{if } \phi_2 \end{cases} \times \begin{cases} g_1 & \text{if } \psi_1 \\ g_2 & \text{if } \psi_n \end{cases} = \begin{cases} f_1 \times g_1 & \text{if } \phi_1, \psi_1 \\ f_1 \times g_2 & \text{if } \phi_1, \psi_2 \\ f_2 \times g_1 & \text{if } \phi_2, \psi_1 \\ f_2 \times g_2 & \text{if } \phi_2, \psi_2 \end{cases} \tag{10}$$

$$f|_{x \leftarrow \frac{F}{G}} = \begin{cases} f_1|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_1|_{x \leftarrow \frac{F}{G}} \\ \vdots & \\ f_m|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_m|_{x \leftarrow \frac{F}{G}} \end{cases} \tag{11}$$

They are also closed under polynomial fractional substitution (11). The reason is that in the r.h.s of (11), 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 (12) shows, they can be restated as (multiple) case-statements with polynomial conditions.

$$\left(\begin{cases} f_1 & \text{if } \frac{H_1}{H_2} > 0 \\ \vdots & \end{cases}\right) = \begin{cases} f_1 & \text{if } H_1 > 0, H_2 > 0 \\ f_1 & \text{if } H_1 < 0, H_2 < 0 \\ \ldots & \end{cases} \tag{12}$$

$$\left|\left(\begin{cases} \frac{N_1}{D_1} & \text{if } \phi_1 \\ \vdots & \end{cases}\right)\right| = \begin{cases} \frac{N_1}{D_1} & \text{if } N_1 > 0, D_1 > 0, \phi_1 \\ \frac{N_1}{D_1} & \text{if } N_1 < 0, D_1 < 0, \phi_1 \\ \frac{-N_1}{D_1} & \text{if } N_1 > 0, D_1 < 0, \phi_1 \\ \frac{-N_1}{D_1} & \text{if } N_1 < 0, D_1 > 0, \phi_1 \\ \ldots & \end{cases} \tag{13}$$

Finally, their closure under *absolute value*, as seen in (13), means PPFs are closed under all operations utilized in Theorem 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:

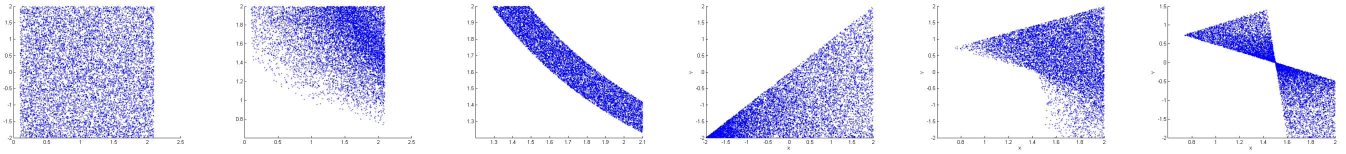
1. Each atomic constraint $\varphi_{i,j}$ can be factorized into a product of some terms in which the maximum degree of each variable is less or equal to 2.
2. The denominator of each sub-function can be factorized into polynomials in which the maximum degree of each variable s less or equal to 2.

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

$$\frac{x^2y^3 + 7xz + 10}{(5xy^2 + 2)(y + x)^3} \quad \text{if } (y^2 + z^2 - 1)(x^2 + 2xy) > 0 \tag{14}$$

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

[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.



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

Figure 2: 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.

instantiated, the resulting univariate function has a closed form integral. This is sufficient for Gibbs sampling since in each step, only one variable is uninstantiated. However, we want to go a step further and compute univariate integrals of multivariate piecewise functions *without* instantiating the remaining variables to avoid the need for an integration per sample. 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.

The following procedure computes $\int_\alpha^\beta f \, \mathrm{d}x$ 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, (14) can be restated as:

$$\begin{cases} \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1)>0, (x^2+2xy)>0 \\ \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1)<0, (x^2+2xy)<0 \end{cases} \tag{15}$$

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 (16) is replaced by cases statements (17).

$$f_1 \quad \text{if } (A \cdot x^2 + B \cdot x + C) > 0 \tag{16}$$

$$\begin{cases} f_1 & \text{if } (A>0), \left(x > \frac{-B+\sqrt{B^2-4AC}}{2A}\right) \\ f_1 & \text{if } (A>0), \left(x < \frac{-B-\sqrt{B^2-4AC}}{2A}\right) \\ f_1 & \text{if } (A<0), \left(x > \frac{-B-\sqrt{B^2-4AC}}{2A}\right), \left(x < \frac{-B+\sqrt{B^2-4AC}}{2A}\right) \end{cases} \tag{17}$$

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:

$$\int_\alpha^\beta \Big[x^3 + xy \quad \text{if } (x>\mathbf{3}), (x>\mathbf{y}), (x<\mathbf{y^2-7}), (\mathbf{y>0}) \Big] \, \mathrm{d}x$$

$$= \Big[\int_{\max\{\alpha,\, \mathbf{3},\, \mathbf{y}\}}^{\min\{\beta,\, \mathbf{y^2-7}\}} x^3 + xy \, \mathrm{d}x \Big] \quad \text{if } (\mathbf{y>0})$$

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.

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.

$$\mathcal{F}\colon \{X_1, \ldots X_N\} \to (\mathbb{R}^N \to \mathbb{R}^+ \cup \{0\})$$

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

Note that the difference between (4) and (18) is that in the former, all variables except X_i are already instantiated therefore 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. 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 we show experimentally in the next section.

Experimental Results

In this section, we are interested in (a) comparing the efficiency and accuracy of our proposed *Symbolic Gibbs* against other MCMC methods on graphical models with piecewise factors and observed determinism, as well as (b) studying the performance of collapsing determinism (as we propose) vs. the practice of relaxing such constraints with noise (as often required in probabilistic programming toolkits).

Algorithms compared. We compare the proposed *Symbolic Gibbs* (SymGibbs) sampler 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 3 depicts 10000 samples are taken from the posterior of Figure 2-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) (19) 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).

$$\mathrm{MAE}_\Gamma(t) := \frac{1}{\zeta \cdot n_\Gamma^t} \sum_{j=1}^{\zeta} \sum_{i=1}^{n_\Gamma^t} \left| q_{j,\Gamma}^{(i)} - q_j^* \right| \tag{19}$$

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.

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:

$$M^* = V^* = \sqrt{P_{\text{tot}}/n} \tag{20}$$

Conditioned on $P_{\text{tot}} = 1.5n$, all masses M_i and velocities V_i are queried. By (20), all elements of the ground truth vector $\mathbf{q}^*$ are $\sqrt{1.5}$. MAE vs. time is depicted in Figures 4.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 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:

$$\frac{1}{R_1} + \ldots + \frac{1}{R_n} = c \tag{21}$$

where $c = \frac{I}{V}$. Equations if the form (21) 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 5.a & b respectively.

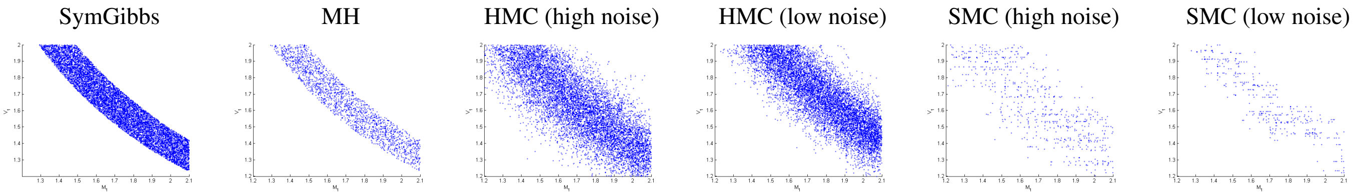
[6] We also tested the other algorithm implemented by Anglican, namely *Particle-Gibbs* (PGibbs) (a variation of Particle-MCMC(ADH10)) and *random database* (RDB) (an MH-based algorithm introduced in (WSG11)) (see (WvdMM14)). In our experimental models, the performance of these algorithms is very similar 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.

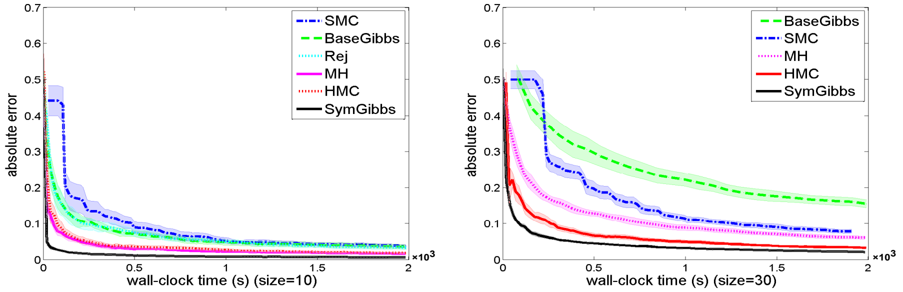
Table 1: Parameters corresponding each experimental model

#	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	$\sigma^2_G = 0.02$	$\sigma^2_G = 0.07$	$G = n/10.17$



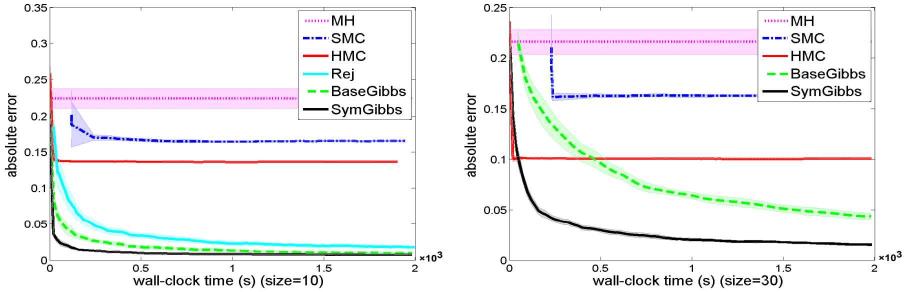
(a) (b) (c) (d) (e) (f)

Figure 3: 10000 samples taken from the distribution Fig. (2-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.



(a) (b)

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



(a) (b)

Figure 5: 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.

Experimental evaluations. Plots of Figure 3 shows that MH and SMC suffer from low *effective sample size*. Note that the apparent sparsity of plots 3-b, 3-e & 3-f is due to repeated samples (rejected proposals). The carried out quantitative measurements (Figures 4 and 5) indicate that in all experimental settings, *Symbolic Gibbs* constantly and significantly performs the best.

All quantitative measurements (Figures 4 and 5) 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.

Conclusion

In this paper, we introduced an expressive class of piecewise algebraic graphical models with nonlinear determinism using factors specified in a rich class of polynomial piecewise fractional functions (PPFs) using polynomial partitioning constraints. We showed that a large subset of PPFs have symbolic univariate integrals, which together with collapsing of nonlinear determinism, 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 offline. Hence, *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 purview of existing inference methods.

References

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.

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

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.

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

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

Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

Andrei Nikolaevich Kolmogorov. Foundations of the theory of probability. 1950.

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.

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.

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

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

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

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

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.

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

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.

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.