Closed-Form Gibbs Sampling for Graphical Models with Algebraic Constraints

Hadi Mohasel Afshar
Research School of Computer Science
Australian National University
Canberra, ACT 0200, Australia
hadi.afshar@anu.edu.au

Scott Sanner
School of EE & Computer Science
Oregon State University
Corvallis, OR 973331, USA
scott.sanner@oregonstate.edu

Christfried Webers
National ICT Australia (NICTA)
Canberra, ACT 2601, Australia
christfried.webers
@nicta.com.au

Abstract

Probabilistic inference in many real-world problems requires graphical models with deterministic algebraic constraints between random variables (e.g., Newtonian mechanics, Pascal's law, Ohm's law) that are known to be problematic for many inference methods such as Monte Carlo sampling. Fortunately, when such constraints are invertible, the model can be collapsed and the constraints eliminated through the well-known Jacobian-based change of variables. As our first contribution in this work, we show that a much broader class of algebraic constraints can be collapsed by leveraging the properties of a Dirac delta model of deterministic constraints. Unfortunately, the collapsing process can lead to highly piecewise densities that pose challenges for existing probabilistic inference tools. Thus, our second contribution to address these challenges is to present a variation of Gibbs sampling that efficiently samples from these piecewise densities. The key insight to achieve this is to introduce a class of functions that (1) is sufficiently rich to approximate arbitrary models up to arbitrary precision, (2) is closed under dimension reduction (collapsing) for models with (non)linear algebraic constraints and (3) always permits one analytical integral sufficient to automatically derive closed-form conditionals for Gibbs sampling. Experiments demonstrate the proposed sampler converges at least an order of magnitude faster than existing Monte Carlo samplers.

Introduction

Probabilistic inference in many real-world problems requires graphical models with deterministic algebraic constraints between random variables. Consider the following running example from physics and the associated 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). Letting $\mathcal{U}(a,\, b)$ denote a uniform density with support $[a, b]$, the prior density of masses and velocities are:*

$$p(M_1) = \mathcal{U}(0.1,\, 2.1), \qquad p(M_2) = \mathcal{U}(0.1,\, 2.1) \quad (1)$$

$$p(V_1) = \mathcal{U}(-2,\, 2), \qquad p(V_2 \mid V_1) = \mathcal{U}(-2,\, V_1) \quad (2)$$

Total momentum is observed to be 3.0 *yielding constraints:*

$$P_1 = M_1V_1, \quad P_2 = M_2V_2, \quad P_{tot} = P_1 + P_2 = 3.0 \quad (3)$$

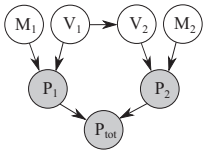


Figure 1: Bayes net for the *collision model*. Shaded circles correspond to random variables that are functions of other variables.

In such problems, the posterior densities only have support on (non)linear sub-manifolds of the parameter space (e.g. the manifold induced by (3) in the collision model). Efficient inference on such models is challenging (Pennec 2006). To evade these complications, state-of-the-art MCMC based probabilistic inference tools suggest adding noise to the deterministic constraints.[1] Unfortunately, this strategy can be problematic: 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 (Li, Ramsundar, and Russell 2013; Chin and Cooper 1987).

The other potential solution is to reduce the dimensionality of the posterior via Jacobian-based variable transformations. Measure theoretic subtleties aside, such transformations are only applicable when the deterministic constraint is invertible with respect to at least one variable. Using the properties of the Dirac delta, our first contribution is to propose a *dimension reduction* (or *collapsing*) method that is more general in the sense that the constraint is not

[1] The state-of-the-art probabilistic programming languages, disallow deterministic continuous random variables be observed. For instance, in BUGS (Lunn et al. 2009), *logical nodes* cannot be given data or initial values. In PyMC (Patil, Huard, and Fonnesbeck 2010) deterministic variables have no *observed flag*. In Stan (Stan Development Team 2014) 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 (Wood, van de Meent, and Mansinghka 2014) throws error "invalid-observe", etc. Therefore, they cannot handle observed constraints natively except by adding noise to the observation. E.g. (in collision model) approximating $P_{\text{tot}} = 3$ with a normal distribution $\mathcal{N}(P_{\text{tot}} - 3, \sigma_\eta^2)$ where the variance σ_η^2 is the noise parameter.

required to be invertible but should be solvable with one or several distinct roots. To our knowledge, this is the first time that Dirac delta constraints for non-invertible functions have been shown to yield collapsible graphical models w.r.t. these constraints. Nonetheless, dimension reduction (either carried out via Jacobian-based or Dirac delta-based mechanism) does not fully eliminate inferential difficulties since as it will be shown shortly, the produced low-dimensional densities are highly piecewise and multimodal. Inference for such *collapsed models* can be extremely challenging.

To date, applicable exact inference tools for piecewise distributions are restricted to piecewise polynomial models where the partitioning boundaries are respectively hyperrectangular, hyper-rhombus or linear (Shenoy and West 2011; Shenoy 2012; Sanner and Abbasnejad 2012) and the (collapsed) observed constraints are restricted to linear equations. To handle a nonlinear observed constraint, (Cobb and Shenoy 2005) approximate it by several piecewise linear constraints by dividing the space into hypercubes. This cannot be used in high-dimensional approximations since the number of partitions required to preserve reasonable accuracy is exponential in dimensionality. Furthermore, constraints aside, exact inference in models with piecewise distributions can easily be intractable since the number of posterior partitions may grow exponentially in the number of marginalizations (required in the forthcoming (5)).

As an alternative to exact inference, asymptotically unbiased approximate inference methods like Monte Carlo sampling can be used. Nonetheless, convergence of most of these algorithms do not hold for the aforementioned piecewise collapsed models, hence resulting in poor convergence rates as our experiments will show. For instance, the leapfrog mechanism by which Hamiltonian Monte Carlo (HMC) simulates the Hamiltonian dynamics relies on the assumption of smoothness (Neal 2011). This assumption does not hold in the adjacency of discontinuities (borders of pieces) leading to low proposal acceptance rates and poor performance. Slice sampling (Neal 2003) suffers from the multimodal nature of the distributions that arise in this 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 density* and the suddenly varying target density 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 density and therefore follow the target changes and multimodalities. Nonetheless, Gibbs samplers can be quite slow since the per sample computation of conditional CDFs that Gibbs relies on are costly and in general the required integral cannot be performed in closed-form.

After presenting collapsing of (nonlinear) algebraic constraints in the first part of this paper, in the second part we address the problem of sampling from the resulting highly piecewise densities. To do this, we introduce a rich 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 dimension reduction of constraints expressed as Dirac deltas. This class is rich enough to approximate arbitrary density functions up to arbitrary precision. We further show that the form of the 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 online sampling computation. 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 (GM). Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ be a set of random variables with realizations in the form $\mathbf{x} = \{x_1, \ldots, x_N\}$.[2] For the sake of notational consistency, throughout we assume $\mathbf{X}$ only contain continuous variables. To cover both directed and undirected GMs we use *factor graph* notation (Kschischang, Frey, and Loeliger 2001) and represent a joint probability density $p(\mathbf{X})$ in a factorized form (4) in which Ψ_k are non-negative *potential functions* of subsets $\mathbf{X}_k$ of $\mathbf{X}$.

$$p(\mathbf{X}) \propto \prod_{\Psi_k \in \mathbf{\Psi}} \Psi_k(\mathbf{X}_k) \tag{4}$$

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*) by (5) where $\mathbf{W} := \mathbf{X} \backslash (\mathbf{Q} \cup \mathbf{E})$ are *marginalized out*.

$$p(\mathbf{Q} \,|\, \mathbf{E} = \mathbf{e}) \propto \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\mathbf{Q}, \mathbf{W} = \mathbf{w}, \mathbf{E} = \mathbf{e}) \, \mathrm{d}\mathbf{w} \tag{5}$$

The integrals required in (5) are often intractable and hence we must often resort to MCMC methods such as Gibbs sampling (Geman and Geman 1984) — 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 (6) and samples are taken via inverse transform sampling.

$$\mathrm{CDF}(X_i \,|\, \mathbf{x}_{-i}) \propto \int_{-\infty}^{X_i} p(X_i = t, \, \mathbf{X}_{-i} = \mathbf{x}_{-i}) \, \mathrm{d}t \tag{6}$$

Collapsing Observed Constraints

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

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

[3] This is to prevent the Borel-Kolmogorov paradox (Kolmogorov 1950) that arises when conditioning on an event with a

In the following theorem, we use the calculus of Dirac deltas and generalize the concept of change of random variables to (not necessarily) invertible functions $f(x_1, \cdot)$. Since in formula (7) one variable is collapsed (i.e., marginalized out), we refer to it as *dimension reduction*.

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

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

where $f(x_1,\ldots,x_n) - z = 0$ has real and simple roots for x_1 with a non-vanishing continuous derivative $\partial f(x_1,\ldots,x_n)/\partial x_1$ at all those roots. Denote the set of all roots by $\mathcal{X}_1 = \{x_1 \mid f(x_1,\ldots,x_n) - z = 0\}$. (Note that each element of $\mathcal{X}_1$ is a function of the remaining variables $x_2,\ldots,x_n,z$.) Then:

$$p(x_2,\ldots,x_n|Z{=}z) \propto \sum_{x_1^i \in \mathcal{X}_1} \frac{p(X_1 = x_1^i, x_2,\ldots,x_n)}{\Big|\big(\partial f(x_1,\ldots,x_n)/\partial x_1\big)|_{x_1 \leftarrow x_1^i}\Big|} \tag{7}$$

Proof. $p(x_2,\ldots,x_n \mid Z = z) \propto$

$$\int_{-\infty}^{\infty} p(x_1,\ldots,x_n) p(Z = z \mid x_1,\ldots,x_n)\, \mathrm{d}x_1 = \int_{-\infty}^{\infty} p(x_1,\ldots,x_n)\delta\big(f(x_1,\ldots,x_n) - z\big)\, \mathrm{d}x_1 \tag{8}$$

According to (Gel'fand and Shilov 1964) there is a unique way to define the composition of Dirac delta with an arbitrary function $h(x)$:

$$\delta(h(x)) = \sum_i \frac{\delta(x - r_i)}{|\partial h(x)/\partial x|} \tag{9}$$

where r_i are all (real and simple) roots of $h(x)$ and $h(x)$ is continuous and differentiable in the root points. By (8), (9) and *Tonelli's theorem*[4] $p(x_2,\ldots,x_n \mid Z = z) \propto$

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

which implies (7). □

To clarify the theorem, it is used to compute the collision model posterior $p(M_2, V_1, V_2 \mid P_{\text{tot}} = 3)$ as follows:

In this model the prior joint density $p(M_1, M_2, V_1, V_2)$ is the product of potentials in equations (1) and (2) which is,

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

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 $(3 - M_2V_2)/V_1$).[5] *Since,*

$$\left|\frac{\partial(M_1V_1 + M_2V_2)}{\partial M_1}\right| = |V_1|$$

by (7), $p(M_2, V_1, V_2 \mid P_{tot} = 3)$ is proportional to

$$\begin{cases} \frac{1}{V_1(16V_1+32)} & \text{if } 0<V_1,\, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ \frac{-1}{V_1(16V_1+32)} & \text{if } V_1<0,\, 0.1<\frac{3-M_2V_2}{V_1}<2.1, \\ & 0.1<M_2<2.1, -2<V_1<2, -2<V_2<V_1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Using (10), various queries are evaluated and depicted in Figure 2. These plots clearly illustrate that even in this low-dimensional example, reducing the dimensionality can lead to multimodal and piecewise posteriors that do not resemble the smooth densities often studied in the literature.

In the next section, we introduce a class of piecewise polynomial functions which is closed under dimension reduction for *algebraic constraints* (e.g., polynomials) and consequently suitable for use with algebraically constrained models. The only task that remains then is to provide an automated sampling method for such models which will be presented subsequently.

Polynomial Piecewise Fractionals (PPFs)

We introduce an expressive family of functions that is rich enough to simulate arbitrary density functions up to arbitrary precision. This family is the class of *polynomial piecewise fractional* functions (PPFs). More formally, 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,

$$f = \begin{cases} f_1 & \text{if } \phi_1 \\ \vdots & \\ f_m & \text{if } \phi_m \end{cases} = \begin{cases} \frac{N_1}{D_1} & \text{if } \varphi_{1,1} \lesseqgtr 0,\, \varphi_{1,2} \lesseqgtr 0,\, \ldots \\ \vdots & \\ \frac{N_m}{D_m} & \text{if } \varphi_{m,1} \lesseqgtr 0,\, \varphi_{m,2} \lesseqgtr 0,\, \ldots \end{cases} \tag{11}$$

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 $<$)[6] where each *atomic constraint* $\varphi_{i,j}$ is a polynomial.

An important property of the class of PPFs is that it is closed under operations required in (7). This paves the way for automated (and potentially multiple) applications of Theorem 1. To show this, note that by (12), PPFs are closed under elementary operations.

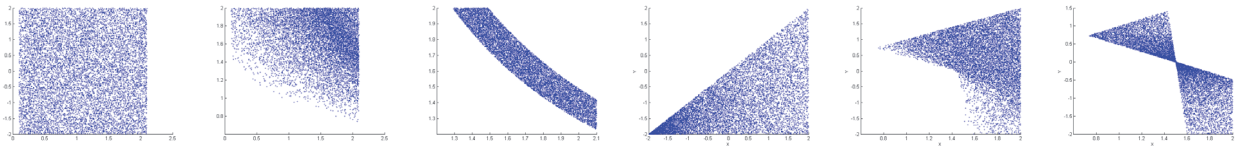
$$\begin{cases} f_1 & \text{if } \phi_1 \\ f_2 & \text{if } \phi_2 \end{cases} \otimes \begin{cases} g_1 & \text{if } \psi_1 \\ g_2 & \text{if } \psi_n \end{cases} = \begin{cases} f_1 \times g_1 & \text{if } \phi_1, \psi_1 \\ f_1 \times g_2 & \text{if } \phi_1, \psi_2 \\ f_2 \times g_1 & \text{if } \phi_2, \psi_1 \\ f_2 \times g_2 & \text{if } \phi_2, \psi_2 \end{cases} \tag{12}$$

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.

[5] In this example, the constraint has a single root (therefore invertible) but if it had several roots, the theorem could still be applied in a straightforward way. As an alternate example, consider $\delta(z - f(x_1, x_2))$ with $z = 0$ and $f(x_1, x_2) = (x_1 - x_2)(x_1 + x_2)$, which yields two roots where the Jacobian cannot be applied, but Theorem 1 can be applied.

[6] We assume the total measure on the border of partitions is 0.



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

Figure 2: Prior/posterior joint density functions 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 model generated dimension reduction algorithm (equation 10).

$$f|_{x \leftarrow \frac{F}{G}} = \begin{cases} f_1|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_1|_{x \leftarrow \frac{F}{G}} \\ \vdots & \\ f_m|_{x \leftarrow \frac{F}{G}} & \text{if } \phi_m|_{x \leftarrow \frac{F}{G}} \end{cases} \tag{13}$$

They are also closed under polynomial fractional substitution (13). The reason is that firstly, sub-functions $f_i|_{x \leftarrow \frac{F}{G}}$ are polynomial fractions and secondly, although conditions $\phi_i|_{x \leftarrow \frac{F}{G}}$ are fractional, as (14) shows, they can be restated as (multiple) case-statements with polynomial conditions.

$$\begin{pmatrix} f_1 & \text{if } \frac{H_1}{H_2} > 0 \\ \vdots & \end{pmatrix} = \begin{cases} f_1 & \text{if } H_1 > 0, H_2 > 0 \\ f_1 & \text{if } H_1 < 0, H_2 < 0 \\ \dots & \end{cases} \tag{14}$$

Similarly, PPFs are closed under *absolute value*.

Analytic integration

Large subsets of PPF have closed-form single integrals. In the next section, we propose a sampling method that performs significantly well on such subsets. For simplicity we only focus on the following form:

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

1. Atomic constraints $\varphi_{i,j}$ are factorized into terms where the maximum degree of each variable is at most 2.
2. The denominator of each sub-function can be factorized into polynomials in which the maximum degree of each variable is at most 2.

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

$$\frac{x^2y^3 + 7xz + 10}{(5xy^2 + 2)(y + x)^3} \quad \text{if } (y^2 + z^2 - 1)(x^2 + 2xy) > 0 \tag{15}$$

Note that by (4), GMs are often designed in factorized forms, therefore, the verification of PPF* conditions is often not hard.

Analytic univariate PPF* integration. Now we provide a procedure for analytic 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. This is sufficient for *exact* 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, (15) can be restated as:
$$\begin{cases} \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1) > 0, (x^2+2xy) > 0 \\ \frac{x^2y^3+7xz+10}{(5xy^2+2)(y+x)^3} & \text{if } (y^2+z^2-1) < 0, (x^2+2xy) < 0 \end{cases} \tag{16}$$
3. *(Isolation of integrand).* 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 (17) is replaced by (18).
$$f_1 \quad \text{if } (A \cdot x^2 + B \cdot x + C) > 0 \tag{17}$$
$$\begin{cases} f_1 & \text{if } (A>0), (x > \frac{-B+\sqrt{B^2-4AC}}{2A}) \\ f_1 & \text{if } (A>0), (x < \frac{-B-\sqrt{B^2-4AC}}{2A}) \\ f_1 & \text{if } (A<0), (x > \frac{-B-\sqrt{B^2-4AC}}{2A}), (x < \frac{-B+\sqrt{B^2-4AC}}{2A}) \end{cases} \tag{18}$$

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:

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

Table 1: Parameters of each experimental model

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

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 via polynomial division (in case the degree of x in the sub-function's numerator is more than its denominator), followed by partial fraction decomposition.

Closed-form Gibbs Sampling

Closed-form Gibbs sampling is based on a simple but significantly useful insight: If $p(\mathbf{X})$ has analytical integrals w.r.t. any 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.

$$\mathcal{F}\colon \{X_1, \dots X_N\} \to (\mathbb{R}^N \to \mathbb{R}^+ \cup \{0\})$$
$$X_i \mapsto \int_{-\infty}^{X_i} p(X_i = t, \mathbf{X}_{-i})\, \mathrm{d}t \qquad (19)$$

Note that the difference between (6) and (19) is that in the former, all variables except X_i are already instantiated therefore $\mathrm{CDF}(X_i \,|\, \mathbf{x}_{-i})$ 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 not costly. 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 *closed-form Gibbs* against other MCMC methods on models with observed constraints as well as (b) studying the performance of the proposed *collapsing* mechanism (*dimension reduction*) vs. the practice of relaxing such constraints with noise (as often suggested in probabilistic programming toolkits).

Algorithms compared

To address item (a), we compare the proposed *closed-form Gibbs* sampler (SymGibbs) to *baseline Gibbs* (BaseGibbs) (Pearl 1987), *rejection sampling* (Rej), *tuned Metropolis-Hastings* (MH)[7] and *Hamiltonian Monte Carlo* (HMC) (Neal 2011) on the collapsed models using the Stan probabilistic programming language (Stan Development Team 2014) and the Anglican implementation of *Sequential Monte Carlo* (SMC) using the Anglican probabilistic programming language (Wood, van de Meent, and Mansinghka 2014). SymGibbs and BaseGibbs require no tuning. MH is automatically tuned after (Roberts et al. 1997) by testing 200 equidistant proposal variances in interval $(0, 0.1]$ and accepting a variance for which the acceptance rate closer to 0.24. HMC on collapsed models produces results similar to MH (in high dimensional piecewise models, both methods reject almost every proposal). Therefore, the results of the former algorithm are not depicted for the readability of the plots.

To answer item (b), HMC and SMC on the models where noise is added to the observations are plotted. To soften the determinism, 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^2_Z. Anglican's syntax requires adding noise to all observed variables. Therefore, in the case of SMC, stochastic observations are also associated with noise parameters. The used parameters are summarized in Table 1.

We also tested *Particle-Gibbs* (PGibbs) (a variation of Particle-MCMC (Andrieu, Doucet, and Holenstein 2010)) and *random database* (RDB) (an MH-based algorithm introduced in (Wingate, Stuhlmueller, and Goodman 2011)). In our experimental models, the performance of these algorithms is very similar to (SMC). Therefore, for readability of the plots, they are not depicted. All algorithms run on a 4 core, 3.40GHz PC.

Measurements

To have an intuitive sense of the performance of different MCMCs, Figure 3 depicts 10000 samples that 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 a query vector $\mathbf{Q} = [Q_1, \dots, Q_\zeta]$. The number of samples taken by a Markov chain Γ up to a time t is denoted by n^t_Γ and the samples are denoted by $\mathbf{q}^{(1)}_\Gamma, \dots, \mathbf{q}^{(n^t_\Gamma)}_\Gamma$ where $\mathbf{q}^{(i)}_\Gamma := [q^{(i)}_{1,\Gamma}, \dots, q^{(i)}_{\zeta,\Gamma}]$. We measure mean absolute error (MAE) of equation (20) vs (wall-clock) time t where $\mathbf{q}^* := [q^*_1, \dots q^*_\zeta]$ is the ground truth mean query vector (that is computed manually due to the symmetry of the models).

$$\mathrm{MAE}_\Gamma(t) := \frac{1}{\zeta \cdot n^t_\Gamma} \sum_{j=1}^{\zeta} \sum_{i=1}^{n^t_\Gamma} \left| q^{(i)}_{j,\Gamma} - q^*_j \right| \qquad (20)$$

[7] MH is automatically tuned after (Roberts et al. 1997) by testing 200 equidistant proposal variances in interval $(0, 0.1]$ and accepting a variance for which the acceptance rate closer to 0.24.

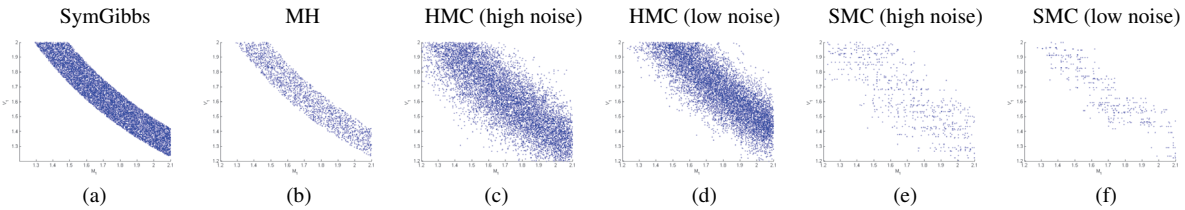


Figure 3: 10000 samples taken from the density of Figure (2-c) using (a) *closed-form Gibbs* sampler and (b) MH with *proposal variance* 0.8 on the reduced-dimension model as well as HMC with (c) measurement error variance 0.2 and (d) 0.01 as well as SMC with parameters (e) $\sigma^2_{V_2} = 0.01$, $\sigma^2_{P_{\text{tot}}} = 0.2$ and (f) $\sigma^2_{V_2} = 0.01$, $\sigma^2_{P_{\text{tot}}} = 0.1$.

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

Experimental models

Although PPF*s are rich enough to approximate arbitrary models, the approximation mechanism is beyond the scope of the present work. As a result we choose experimental models that are already in such algebraic forms.

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:

$$M^* = V^* = \sqrt{P_{\text{tot}}/n} \tag{21}$$

Conditioned on $P_{\text{tot}} = 1.5n$, all masses M_i and velocities V_i are queried. By (21), all elements of the ground truth vector $\mathbf{q}^*$ are $\sqrt{1.5}$.

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

$$\frac{1}{R_1} + \ldots + \frac{1}{R_n} = c \tag{22}$$

where $c = \frac{I}{V}$. Equations if the form (22) 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:

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

Figure 4 depicts MAE vs. time plots associated with the introduced models of different sizes.

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 and 3-f is due to repeated samples (rejected proposals). The carried out quantitative measurements (Figure 4) indicate that in all experimental settings, *closed-form Gibbs* consistently performs the best while its superiority in high dimensions is significant.

Particularly in the Building wiring model (which is more complicated and highly piecewise), the quantitative measurements indicate that hard to soft constraint conversion (via introducing noise for measurement error) ends in poor results (Figures 4-c and 4-d). Interestingly, in this model, even in a dimensionality as low as 10, the Metropolis-Hasting based algorithms (i.e., MH, 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 presented a mechanism to carry out probabilistic inference conditioned on observed algebraic constraints, i.e., algebraic functions of continuous random variables, via a *collapsing mechanism* to reduce the dimensionality of the variable space. The proposed method is based on the properties of the Dirac delta and is more general than Jacobian-based change of variables in the sense that it does not require the observed functions to be invertible w.r.t. any variable.

Nonetheless, dimension reduction often leads to highly piecewise and multimodal posteriors. This is a bridge to the second part of the paper where we show that on an expressive family of models, the costly operations required for Gibbs sampling can be performed analytically and prior to the sampling process. This leads to an automated and closed-form sampler that is significantly faster than the baseline. The studied family of models is rich enough to express algebraic constraints as well as being able to approximate arbitrary density functions up to arbitrary precisions. Our experimental results show that (1) the alternative to dimension reduction, i.e., adding noise to the observations leads to unsatisfactory results; and (2) on the piecewise posteriors generated via the collapsing mechanism, the performance of the proposed closed-form Gibbs sampler can be overwhelmingly superior to the other samplers executed on the same model.

The combination of these novel contributions makes probabilistic reasoning applicable to variety of new applications

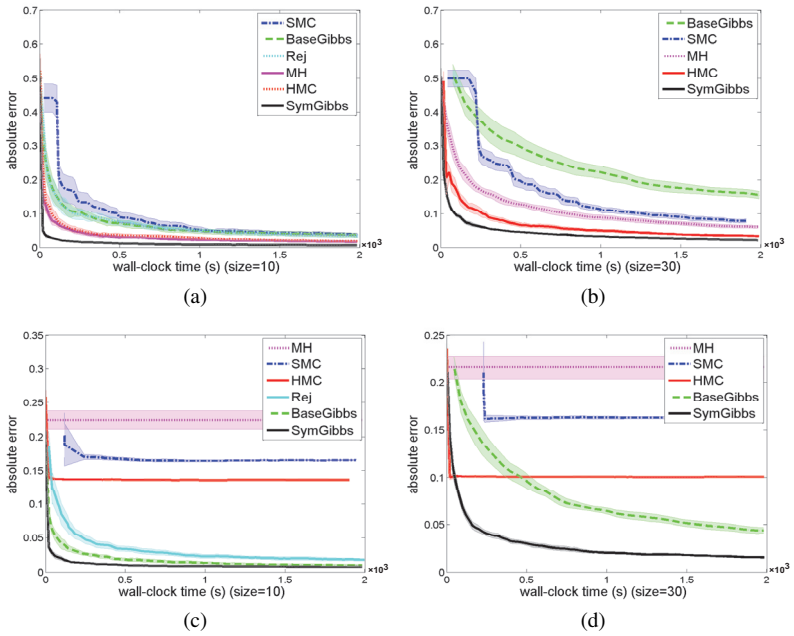


Figure 4: Mean absolute error versus time in Multi-object collision model with (a) 10 and (b) 30 objects as well as Building wiring model with (c) 10 and (d) 30 paralleled resistors.

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

References

Andrieu, C.; Doucet, A.; and Holenstein, R. 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3):269–342.

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

Cobb, B. R., and Shenoy, P. P. 2005. Nonlinear deterministic relationships in Bayesian networks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer. 27–38.

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

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

Kolmogorov, A. N. 1950. Foundations of the theory of probability.

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

Li, L.; Ramsundar, B.; and Russell, S. 2013. Dynamic scaled sampling for deterministic constraints. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*.

Lunn, D.; Spiegelhalter, D.; Thomas, A.; and Best, N. 2009. The BUGS project: Evolution, critique and future directions. *Statistics in medicine* 28(25):3049–3067.

Neal, R. M. 2003. Slice sampling. *Ann. Statist.* 31(3):705–767.

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

Patil, A.; Huard, D.; and Fonnesbeck, C. J. 2010. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software* 35(4):1.

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

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

Roberts, G. O.; Gelman, A.; Gilks, W. R.; et al. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability* 7(1):110–120.

Sanner, S., and Abbasnejad, E. 2012. Symbolic variable elimination for discrete and continuous graphical models. In *AAAI*.

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

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

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

Wingate, D.; Stuhlmueller, A.; and Goodman, N. D. 2011. Lightweight implementations of probabilistic programming languages via transformational compilation. In *International Conference on Artificial Intelligence and Statistics*, 770–778.

Wood, F.; van de Meent, J. W.; and Mansinghka, V. 2014. A new approach to probabilistic programming inference. In *Proceedings of the 17th International conference on Artificial Intelligence and Statistics*.