Fast Bayesian Inference in Piecewise Graphical Models

Abstract

Many real-world Bayesian inference problems such as preference learning or trader valuation modeling in financial markets naturally use piecewise likelihoods. Unfortunately, exact closed-form inference in the underlying Bayesian graphical models is intractable in the general case and existing approximation techniques provide few guarantees on both approximation quality and efficiency. While (Markov Chain) Monte Carlo methods provide an attractive asymptotically unbiased approximation approach, rejection sampling and Metropolis-Hastings both prove inefficient in practice, and analytical derivation of Gibbs samplers require exponential space and time in the amount of data. In this work, we show how to transform problematic piecewise likelihoods into equivalent mixture models and then provide a blocked Gibbs sampling approach for this transformed model that achieves an *exponential-to-linear* reduction in space and time compared to a conventional Gibbs sampler. This enables fast, asymptotically unbiased Bayesian inference in a new expressive class of piecewise graphical models and empirically requires orders of magnitude less time than rejection, Metropolis-Hastings, and conventional Gibbs sampling methods to achieve the same level of accuracy.

Introduction

Many Bayesian inference problems such as preference learning (Guo and Sanner 2010) or trader valuation modeling in financial markets naturally use piecewise likelihoods (Shogren, List, and Hayes 2000), e.g., preferences may induce constraints on possible utility functions while trader transactions constrain possible instrument valuations. To be concrete, consider the following Bayesian approach to preference learning where our objective is to learn a user's weighting of attributes for classes of items (e.g., cars, apartment rentals, movies) given their responses to pairwise comparison queries over those items:

Example 1 (Bayesian pairwise preference learning (BPPL))**.** Suppose each *item* $\mathbf{a}$ is modeled by an N-dimensional real-valued *attribute choice vector* $(\alpha_1, \ldots, \alpha_N)$. The goal is to learn an *attribute weight vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^N$ that describes the utility of each attribute choice from user responses to preference queries. As commonly done in *multi-attribute utility theory* (Keeney and Raiffa 1993), the overall item utility $u(\mathbf{a}|\,\boldsymbol{\theta})$ is decomposed additively over the attribute choices of $\mathbf{a}$:

$$u(\mathbf{a}|\,\boldsymbol{\theta}) = \sum_{j=1}^{N} \theta_j \cdot \alpha_j$$

User responses are in the form of n queries (i.e. observed data points) d_1 to d_n where d_i is a pairwise comparison of some items $\mathbf{a}_i$ and $\mathbf{b}_i$ with the following possible responses:

- $\mathbf{a}_i \succ \mathbf{b}_i$: In the i-th query, the user prefers item $\mathbf{a}_i$ over $\mathbf{b}_i$.
- $\mathbf{a}_i \preceq \mathbf{b}_i$: In the i-th query, the user does not prefer item $\mathbf{a}_i$ over $\mathbf{b}_i$.

It is assumed that with an *elicitation noise* $0 \leq \eta < 0.5$, the item with a greater utility is preferred:

$$pr(\mathbf{a}_i \succ \mathbf{b}_i \mid \boldsymbol{\theta}) = \begin{cases} u(\mathbf{a}_i|\boldsymbol{\theta}) < u(\mathbf{b}_i|\boldsymbol{\theta}) : \eta \\ u(\mathbf{a}_i|\boldsymbol{\theta}) = u(\mathbf{b}_i|\boldsymbol{\theta}) : 0.5 \\ u(\mathbf{a}_i|\boldsymbol{\theta}) > u(\mathbf{b}_i|\boldsymbol{\theta}) : 1 - \eta \end{cases} \quad (1)$$

$$pr(\mathbf{a}_i \preceq \mathbf{b}_i \mid \boldsymbol{\theta}) = 1 - pr(\mathbf{a}_i \succ \mathbf{b}_i \mid \boldsymbol{\theta}) \quad (2)$$

As the graphical model in Figure 1 illustrates, our posterior belief over the user's attribute weights is provided by the standard Bayesian inference expression: $pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n) \propto pr(\boldsymbol{\theta}) \cdot \prod_{i=1}^{n} pr(d_i|\, \boldsymbol{\theta})$. As also evidenced, since the prior and likelihoods are piecewise distributions, the posterior distribution is also piecewise. ◇

Unfortunately, Bayesian inference in models with piecewise likelihoods like BPPL in Example 1 often lead to posterior distributions with a number of piecewise partitions *exponential in the number of data points and attributes*, thus rendering exact analytical inference impossible. While (Markov Chain) Monte Carlo methods provide an attractive asymptotically unbiased approximation approach, rejection sampling and Metropolis-Hastings both prove inefficient in practice, and analytical derivation of Gibbs samplers require exponential space and time in the number of data points and attributes of data.

In this work, we show how to transform problematic piecewise likelihoods into equivalent mixture models and provide a blocked Gibbs sampling approach for this transformed model that achieves an *exponential-to-linear* reduction in space and time compared to a conventional Gibbs sampler. This enables fast, asymptotically unbiased

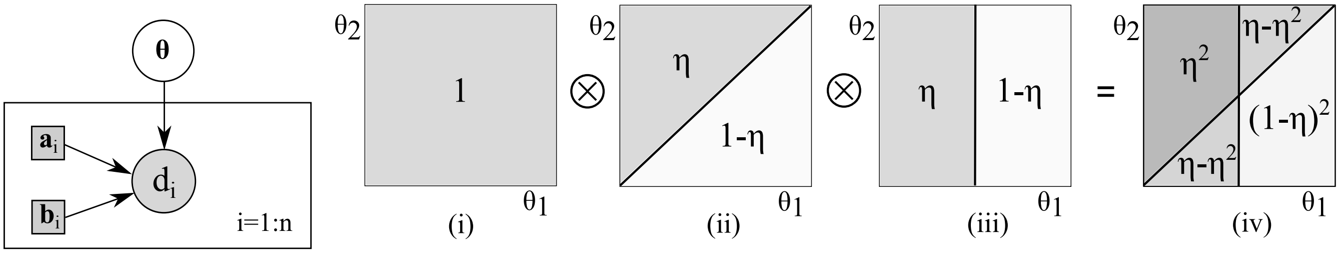


Figure 1: (a) Graphical model for BPPL problem in Example 1. (b) A 2D instance of Example 1: (i) An (unnormalized) prior uniform in a rectangle with center (0,0). (ii) Likelihood model $pr(\mathbf{a}_1 \succ \mathbf{b}_1 | \,\boldsymbol{\theta})$ and (iii) $pr(\mathbf{a}_2 \succ \mathbf{b}_2 | \,\boldsymbol{\theta})$ (as in equation 1) where $\mathbf{a}_1 = (5, 3)$, $\mathbf{b}_1 = (6, 2)$, $\mathbf{a}_2 = \mathbf{a}_1$ and $\mathbf{b}_2 = (6, 3)$. (iv) A piecewise function proportional to the posterior distribution.

Bayesian inference in a new expressive class of piecewise graphical models and empirically requires orders of magnitude less time than rejection, Metropolis-Hastings, and conventional Gibbs sampling methods to achieve the same level of accuracy – especially when the number of posterior partitions grows rapidly in the number of observed data points.

After a brief introduction to piecewise models and the exact/asymptotically unbiased inference methods that can be applied to them in the following section, a novel inference algorithm (referred to as *Augmented Gibbs* sampling throughout) is presented.

Bayesian inference on graphical models with piecewise distributions

Inference. We will present an inference method that can be generalized to a variety of graphical models with piecewise factors, however, our focus in this work is on Bayesian networks factorized in the following standard form

$$pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n) \propto pr(\boldsymbol{\theta},\, d_1, \ldots, d_n) = pr(\boldsymbol{\theta}) \cdot \prod_{i=1}^{n} pr(d_i|\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_N)$ is a parameter vector and d_i are observed data points. A typical inference task with this posterior distribution is to compute the expectation of a function of $f(\boldsymbol{\theta})$ given data:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}) \,|\, d_1, \ldots, d_n\,] \tag{4}$$

Piecewise models. We are interested in the inference on models where prior/likelihood distributions are piecewise. A function $f(\boldsymbol{\theta})$ is m-piece *piecewise* if it can be represented as:

$$f(\boldsymbol{\theta}) = \begin{cases} \phi_1(\boldsymbol{\theta}) : & f_1(\boldsymbol{\theta}) \\ \vdots & \\ \phi_m(\boldsymbol{\theta}) : & f_m(\boldsymbol{\theta}) \end{cases} \tag{5}$$

where ϕ_1 to ϕ_m are mutually exclusive and jointly exhaustive Boolean functions (constraints) that partition the space of variables $\boldsymbol{\theta}$. If for a particular variable assignment $\boldsymbol{\theta}_0$, a constraint $\phi_i(\boldsymbol{\theta}_0)$ is satisfied, then by definition, the function returns the value of its i-th *sub-function*: $f(\boldsymbol{\theta}_0) = f_i(\boldsymbol{\theta}_0)$. In this case, it is said that sub-function f_i is *activated* by assignment $\boldsymbol{\theta}_0$.

In the implementation of our proposed algorithm, the constraints are restricted to linear/quadratic (in)equalities while sub-functions are polynomials with real exponents. However, in theory, the algorithm can be applied to *any* family of piecewise models in which the roots of univariate constraint expressions can be found and sub-functions (and their products) are integrable.

Complexity of inference on piecewise models.

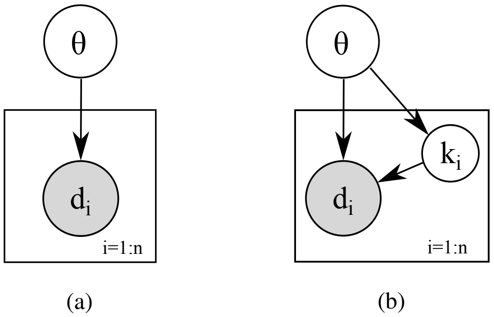
If in the model of equation 3, the prior $pr(\boldsymbol{\theta})$ is an L-piece distribution and each of the n likelihoods is a piecewise function with number of partitions bound by M, then the joint distribution is a piecewise function with number of partitions bound by LM^n (therefore, $O(M^n)$). The reason, as clarified by the following simple formula, is that the number of partitions in the product of two piecewise functions is bound by the product of their number of partitions:[1]

$$\begin{cases} \phi_1(\boldsymbol{\theta}) : f_1(\boldsymbol{\theta}) \\ \phi_2(\boldsymbol{\theta}) : f_2(\boldsymbol{\theta}) \end{cases} \otimes \begin{cases} \psi_1(\boldsymbol{\theta}) : g_1(\boldsymbol{\theta}) \\ \psi_2(\boldsymbol{\theta}) : g_2(\boldsymbol{\theta}) \end{cases} = \begin{cases} \phi_1(\boldsymbol{\theta}) \wedge \psi_1(\boldsymbol{\theta}) : f_1(\boldsymbol{\theta})g_1(\boldsymbol{\theta}) \\ \phi_1(\boldsymbol{\theta}) \wedge \psi_2(\boldsymbol{\theta}) : f_1(\boldsymbol{\theta})g_2(\boldsymbol{\theta}) \\ \phi_2(\boldsymbol{\theta}) \wedge \psi_1(\boldsymbol{\theta}) : f_2(\boldsymbol{\theta})g_1(\boldsymbol{\theta}) \\ \phi_2(\boldsymbol{\theta}) \wedge \psi_2(\boldsymbol{\theta}) : f_2(\boldsymbol{\theta})g_2(\boldsymbol{\theta}) \end{cases} \tag{6}$$

Exact inference on piecewise models. In theory, closed-form inference on piecewise models (at least piecewise polynomials) with linear constraints is possible (Sanner and Abbasnejad 2012). In practice, however, such symbolic methods rapidly become intractable since the posterior requires the representation of $O(M^n)$ distinct case partitions.

Approximate inference on piecewise modes. An alternative option is to seek *asymptotically unbiased* inference methods via Monte Carlo sampling. Given a set of S samples (particles) $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}\}$ taken from a posterior $pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n)$, the inference task of Equation 4 can be approximated by: $\frac{1}{S}\sum_{i=1}^{S} f(\theta^{(i)}|\, d_1, \ldots, d_n)$. Three widely used sampling methods for an arbitrary distribution $pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n)$ are the following:

[1] If pruning potential inconsistent (infeasible) constraint is possible (i.e. by *linear constraint solvers* for linear constrains) and the imposed extra costs are justified, the number of partitions can be less.



(a) (b)

Figure 2: (a) A Bayesian inference model with parameter (vector) $\boldsymbol{\theta}$ and data points d_1 to d_n. (b) A mixture model with parameter (vector) $\boldsymbol{\theta}$ and data points d_1 to d_n

Rejection sampling: Let $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ be two distributions such that direct sampling from them is respectively hard and easy and $p(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ is bound by a constant $c > 1$. To take a sample from p using *rejection sampling*, a sample $\boldsymbol{\theta}$ is taken from distribution q and accepted with probability $p(\boldsymbol{\theta})/cq(\boldsymbol{\theta})$, otherwise it is rejected and the process is repeated. If c is too large, the speed of this algorithm is slow since often lots of samples are required until one is accepted.

Metropolis-Hastings (MH): To generate a new Markov Chain Monte Carlo (MCMC) sample $\boldsymbol{\theta}^{(t)}$ of a distribution $p(\boldsymbol{\theta})$ given a previously taken sample $\boldsymbol{\theta}^{(t-1)}$, firstly, a sample $\boldsymbol{\theta}'$ is taken from a symmetric *proposal density* $q(\boldsymbol{\theta}|\,\boldsymbol{\theta}^{(t-1)})$ (often an isotropic *Gaussian* centered at $\boldsymbol{\theta}^{(t-1)}$). With probability $\min\left(1, q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')/q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})p(\boldsymbol{\theta}^{(t-1)})\right)$, $\boldsymbol{\theta}'$ is accepted $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}'$, otherwise, $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$. Choosing a good *proposal* is problem-dependent and requires tuning. Also most of proposals that often require costly posterior evaluation is rejected leading to a poor performance.

Gibbs sampling: A celebrated MCMC method in which generating new samples $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ requires each variable $\boldsymbol{\theta}_i$ to be sampled conditioned on the last instantiated value of the others: $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i})$. Computation of N univariate *cumulative distribution functions* (CDFs) (one for each $p(\boldsymbol{\theta}_i|\,\boldsymbol{\theta}_{-i})$) as well as their inverse functions is required which can be quite time consuming. In practice, when these integrals are not tractable or easy to compute Gibbs sampling can be prohibitively expensive.

Compared to rejection sampling or MH, the performance of Gibbs sampling on the aforementioned *piecewise* models is exponential in the number of observations and attributes. In this work, we propose an alternative linear Gibbs sampler.

Piecewise models as mixture models

In this section we detail how to overcome the exponential complexity of standard Gibbs sampling by transforming piecewise models to (augmented) mixture models and performing linear time Gibbs sampling in this augmented model.

We motivate the algorithm by first introducing the augmented posterior for piecewise likelihoods:

$$pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n) \;\propto pr(\boldsymbol{\theta}) \otimes \tag{7}$$

$$\begin{cases} \boxed{k_1 = 1.}\, \phi_1^1(\boldsymbol{\theta}) : f_1^1(\boldsymbol{\theta}) \\ \vdots \\ \boxed{k_1 = M.}\, \phi_M^1(\boldsymbol{\theta}) : f_M^1(\boldsymbol{\theta}) \end{cases} \otimes \cdots \otimes \begin{cases} \boxed{k_n = 1.}\, \phi_1^n(\boldsymbol{\theta}) : f_1^n(\boldsymbol{\theta}) \\ \vdots \\ \boxed{k_n = M.}\, \phi_M^n(\boldsymbol{\theta}) : f_M^n(\boldsymbol{\theta}) \end{cases} \tag{8}$$

In the above, k_i is the partition-counter of the i-th likelihood function. ϕ_j^i is its j-th constraint and f_j^i is its associated sub-function. Also for readability, case statements are numbered and without loss of generality, we assume the number of partitions in each likelihood function is M

We observe that each k_i can be seen as a random variable. It deterministically takes the value of the partition whose associated constraint holds (given $\boldsymbol{\theta}$) and its possible outcomes are in $\textsc{Val}(k_i) = \{1, \ldots, M\}$. Note that for any given $\boldsymbol{\theta}$, exactly one constraint holds for a piecewise function, therefore, $\sum_{k_i=1}^{M} pr(k_i\,|\,\boldsymbol{\theta}) = 1$. Intuitively, it can be assumed that k_i is the underlying variable that determines which partition of each likelihood function is 'chosen'. As we have:

$$pr(d_i|\,\boldsymbol{\theta}) = \sum_{k_i=1}^{M} pr(k_i|\,\boldsymbol{\theta}) pr(d_i|\,k_i, \boldsymbol{\theta}), \tag{9}$$

we can claim that a piecewise likelihood function is a mixture model in which sub-function f_j^i is the *mixture component* and k_i provides binary *mixture weight*. Hence:

$$\begin{aligned} pr(\boldsymbol{\theta}|\, d_1, \ldots, d_n) &\propto pr(\boldsymbol{\theta}) \otimes \sum_{k_1} pr(k_1|\,\boldsymbol{\theta}) pr(d_1|\,k_1, \boldsymbol{\theta}) \\ &\qquad \otimes \cdots \otimes \sum_{k_n} pr(k_n|\,\boldsymbol{\theta}) pr(d_n|\,k_n, \boldsymbol{\theta}) \\ &\propto \sum_{k_1} \cdots \sum_{k_n} p(\boldsymbol{\theta}, d_1, \ldots, d_n, k_1, \ldots, k_n) \end{aligned}$$

This means that the Bayesian networks in Figures 2a and 2b are equivalent. Therefore, instead of taking samples from 2a, they can be taken from the *augmented model* 2b. A key observation, however, is that unlike the conditional distributions $pr(\boldsymbol{\theta}_i|\,\boldsymbol{\theta}_{-i})$, in $pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, k_1, \ldots, k_n)$ the number of partitions remains fixed, namely L, rather than growing M^n.

The reason is that if $\mathbf{k} = k_1, \ldots, k_n$ are given, for i-th likelihood, a single sub-function $f_{k_i}^i$ is 'chosen' and $pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, k_1, \ldots, k_n) = pr(\boldsymbol{\theta}|\boldsymbol{\theta}_{-i}) \prod_{k_i} f_{k_i}^i(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i})$. Assuming that sub-functions are not piecewise themselves, the number of partitions in $pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, k_1, \ldots, k_n)$ is bound by the number of partitions in the prior.

In the following proposition we prove that Equation (9) is valid:

Proposition 1. *Piecewise likelihood function* $pr(d|\,\boldsymbol{\theta})$ *defined by (10) is equivalent to* $\sum_{k=1}^{M} pr(k|\,\boldsymbol{\theta}) pr(d|\,k, \boldsymbol{\theta})$

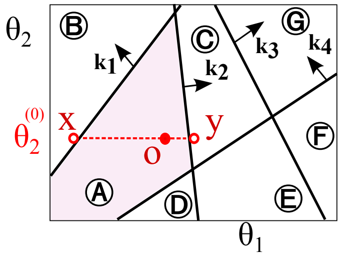


Figure 3: A piecewise joint distribution of (θ_1, θ_2) partitioned by bi-valued linear constraints. In the side, specified by each arrow, its associated auxiliary variable k_i is 1 otherwise 2. A Gibbs sampler started from an initial point $O = (\theta_1^{(0)}, \theta_2^{(0)})$, is trapped in an initial partition (A) where $k_1 = k_2 = k_3 = 1$ and $k_4 = 2$.

where:

$$pr(d|\,\boldsymbol{\theta}) \quad := \quad \begin{cases} \phi_1^d(\boldsymbol{\theta}) & : f_1^d(\boldsymbol{\theta}) \\ & \vdots \\ \phi_M^d(\boldsymbol{\theta}) & : f_M^d(\boldsymbol{\theta}) \end{cases} \tag{10}$$

$$pr(k|\,\boldsymbol{\theta}) \quad := \quad \begin{cases} \phi_k(\boldsymbol{\theta}) & : 1 \\ \neg\phi_k(\boldsymbol{\theta}) & : 0 \end{cases} \tag{11}$$

$$pr(d|\,k, \boldsymbol{\theta}) \quad := \quad f_k^d(\boldsymbol{\theta}) \tag{12}$$

Proof. Since constraints ϕ_k are mutually exclusive and jointly exhaustive:

$$\sum_{k=1}^{N} pr(k|\,\boldsymbol{\theta}) = \sum_{k=1}^{N} \begin{cases} \phi_k(\boldsymbol{\theta}) & : 1 \\ \neg\phi_k(\boldsymbol{\theta}) & : 0 \end{cases} = 1$$

Therefore $pr(k|\,\boldsymbol{\theta})$ is a proper probability function. On the other hand, by marginalizing k, (11) and (12) trivially lead to (10):

$$\begin{aligned} \sum_{k} pr(k|\,\boldsymbol{\theta}) pr(d|\,k, \boldsymbol{\theta}) &= \sum_{k=1}^{N} \begin{cases} \phi_k(\boldsymbol{\theta}) & : 1 \\ \neg\phi_k(\boldsymbol{\theta}) & : 0 \end{cases} \cdot f_k(\boldsymbol{\theta}, d) \quad \text{by (11), (12)} \\ &= \sum_{k=1}^{N} \begin{cases} \phi_k(\boldsymbol{\theta}) & : f_k(\boldsymbol{\theta}, d) \\ \neg\phi_k(\boldsymbol{\theta}) & : 0 \end{cases} \\ &= \begin{cases} \phi_1(\boldsymbol{\theta}) & : f_1(\boldsymbol{\theta}, d) \\ \vdots & \\ \phi_n(\boldsymbol{\theta}) & : f_N(\boldsymbol{\theta}, d) \end{cases} = pr(d|\,\boldsymbol{\theta}) \quad \text{by (10)} \end{aligned}$$

in which the third equality holds since constraints ϕ_k are mutually exclusive. □

Deterministic dependencies and blocked sampling

Deterministic dependencies. It is known that in the presence of determinism Gibbs sampling gives poor results (Poon and Domingos 2006). In our setting, deterministic dependencies arise from Definition (11), were the value of k is decided by $\boldsymbol{\theta}$. This problem is illustrated in Figure 3 by a simple example: A Gibbs sampler started from an initial point $O = (\theta_1^{(0)}, \theta_2^{(0)})$, is trapped in the initial partition (A). The reason is that conditioned on the initial value of the auxiliary variables, the partition is deterministically decided as being (A), and conditioned on any point in (A), the auxiliary variables kept their initial values.

Blocked Gibbs. We avoid deterministic dependencies by Blocked sampling: at each step of Gibbs sampling, a parameter variable $\boldsymbol{\theta}_i$ is jointly sampled with (at least) one auxiliary variable k_j conditioned on the remaining variables: $(\boldsymbol{\theta}_i, k_j) \sim pr(\boldsymbol{\theta}_i, k_j|\,\boldsymbol{\theta}_{-i}, \mathbf{k}_{-j})$. Since $pr(\boldsymbol{\theta}_i, k_j|\,\boldsymbol{\theta}_{-i}, \mathbf{k}_{-j}) = pr(\boldsymbol{\theta}_i|\,\boldsymbol{\theta}_{-i}, \mathbf{k}_{-j}) \cdot pr(k_j|\,\boldsymbol{\theta})$, this is done in 2 steps:

1. k_j is marginalized out and $\boldsymbol{\theta}_i$ is sampled (*collapsed Gibbs sampling*):

$$\boldsymbol{\theta}_i \sim \sum_{k_j} pr(k_j|\,\boldsymbol{\theta}_{-i}, \mathbf{k}_{-j}) pr(\theta_i|\,k_j, \boldsymbol{\theta}_{-i}, \mathbf{k}_{-j})$$

2. k_j is determined from $\boldsymbol{\theta}$: $k_j \leftarrow i \in \textsc{Val}(k_j)$ s.t. $\phi_i^j(\boldsymbol{\theta}) =$ *true* where ϕ_i^j is the i-th constraint of j-th likelihood function.

For instance in Figure 3, for sampling θ_2, either k_1 or k_2 is collapsed, then the next sample will be in the union of partition (A) with either (B) or (C).

Targeted selection of collapsed auxiliary variables. We provide a mechanism for finding auxiliary variables k_j that, with a high probability, are not determined by the other auxiliary variables, $\mathbf{k}_{-j}$ when jointly sampled with a parameter variable $\boldsymbol{\theta}_i$. We observe that the set of partitions satisfying the current valuation of $\mathbf{k}$ often differs with its adjacent partitions in a single auxiliary variable. Since such a variable is not determined by other variables, it can be used in the blocked sampling.

However, in case some likelihood functions share the same constraint, some adjacent partitions would differ in multiple auxiliary variables. In such cases, more than one auxiliary variable should be used in blocked sampling.

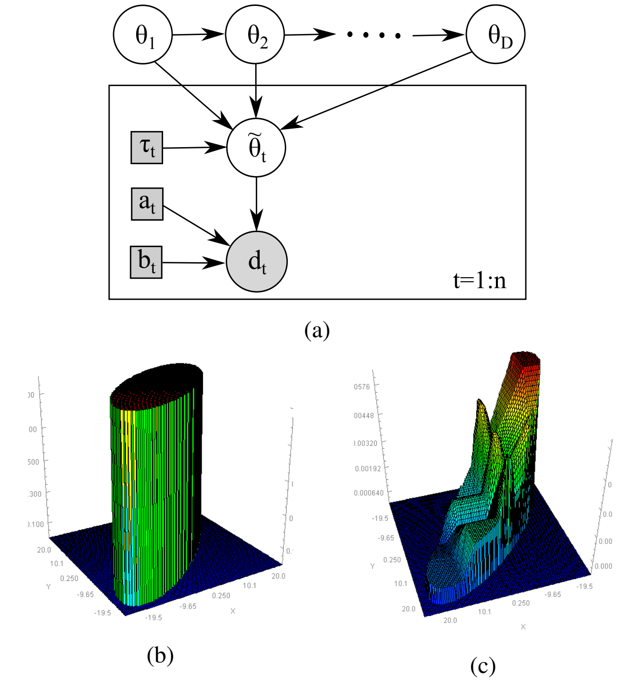
Finding such auxiliary variables has a simple geometric interpretation. As such, instead of trying to formalize the solution, we explain it by a simple example depicted in Figure (3):

Consider finding a proper k_j for blocked sampling $pr(\boldsymbol{\theta}_1, k_j|\,\boldsymbol{\theta}_2^{(0)}, \mathbf{k}_{-j})$. It suffices to extend the line segment $pr(\boldsymbol{\theta}_1|\,\boldsymbol{\theta}_2^{(0)}, \mathbf{k}) > 0$ in two sides to reach the point x or y. In this way, the neighboring partition e.g. (B) and (C) and consequently their corresponding $\mathbf{k}$ valuations are detected. Finally, the auxiliary variables that differ between (A) and (B) or differ between (A) and (C)) are found (k_1 and k_2, in the Figure (3)).

Experimental results

In this section we demonstrate that the mixing time of the proposed method, *augmented Gibbs* sampling, is faster than Rejection sampling, baseline Gibbs and Metropolis-Hastings. Algorithms are tested against two different models. The BPPL model of Example 1 and a *Market maker* (MM) model motivated by (Das and Magdon-Ismail 2008):

Example 2 (Market maker (MM))**.** Suppose there are D different types of *instruments* with respective valuations $\boldsymbol{\theta} =$



(a)

(b)

(c)

Figure 4: Instrument type value distribution of Market Maker problem of Example 2. (a) prior, (b) posterior given 4 observed data points (trader responses).

$\{\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_D\}$. There is a *market maker* who at each time step t deals an instrument of type $\tau_t \in \{1, \ldots D\}$ by setting *bid* and *ask* b_t and a_t denoting prices at which she is willing to buy and sell each unit respectively (where $b_t \leq a_t$). The "true" valuation of different types are unknown (to her) but any a priori knowledge over their dependencies that can be expressed via a DAG structure over their associated random variables is permitted. Nonetheless, without loss of generality we only consider the following simple dependency: Assume the types indicate different versions of the same product and each new version is more expensive than the older ones ($\boldsymbol{\theta}_i \leq \boldsymbol{\theta}_{i+1}$). The valuation of the oldest version is within some given price range $[L, H]$ and the price difference of any consecutive versions is bound by a known parameter δ:

$$
\begin{aligned}
pr(\boldsymbol{\theta}_1) &= \mathcal{U}(L, H) \\
pr(\boldsymbol{\theta}_{i+1}) &= \mathcal{U}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i + \delta) \quad \forall i \in \{1, \ldots, D-1\}
\end{aligned}
$$

At each time-step t, a trader arrives. He has a noisy estimation $\widetilde{\boldsymbol{\theta}}_t$ of the actual value of the presented instrument τ_t. We assume $pr(\widetilde{\boldsymbol{\theta}}_t | \boldsymbol{\theta}, \tau_t = i) = \mathcal{U}(\boldsymbol{\theta}_i - \epsilon, \boldsymbol{\theta}_i + \epsilon)$. The trader response to bid and ask prices a_t and b_t is d_t in $\{\textsc{Buy}, \textsc{Sell}, \textsc{Hold}\}$. If he thinks the instrument is undervalued by the ask price (or overvalued by the bid price), with probability 0.8, he will buy it (resp. sell it), otherwise holds.

$$
pr(\textsc{Buy} \,|\, \widetilde{\boldsymbol{\theta}}_t, a_t, b_t) = \begin{cases} \widetilde{\boldsymbol{\theta}}_t < a_t : 0 \\ \widetilde{\boldsymbol{\theta}}_t \geq a_t : 0.8 \end{cases}
$$

$$
pr(\textsc{Sell} \,|\, \widetilde{\boldsymbol{\theta}}_t, a_t, b_t) = \begin{cases} \widetilde{\boldsymbol{\theta}}_t \leq b_t : 0.8 \\ \widetilde{\boldsymbol{\theta}}_t > b_t : 0 \end{cases}
$$

$$
pr(\textsc{Hold} \,|\, \widetilde{\boldsymbol{\theta}}_t, a_t, b_t) = \begin{cases} b_t < \widetilde{\boldsymbol{\theta}}_t < a_t & : 1 \\ \widetilde{\boldsymbol{\theta}}_t \leq b_t \vee \widetilde{\boldsymbol{\theta}}_t \geq a_t & : 0.2 \end{cases}
$$

Based on traders' responses, the market maker intends to compute the posterior distribution of the valuations of all instrument types. To transform this problem (with corresponding model shown in Figure 4a) to the model represented by Equation 3, variables $\widetilde{\boldsymbol{\theta}}_t$ should be marginalized. For instance:

$$
\begin{aligned}
pr(\textsc{Buy} \,|\, \boldsymbol{\theta}, a_t, b_t, \tau_t) &= \int_{-\infty}^{\infty} pr(\textsc{Buy} \,|\, \widetilde{\boldsymbol{\theta}}_t, a_t, b_t) \cdot pr(\widetilde{\boldsymbol{\theta}}_t \,|\, \boldsymbol{\theta}, \tau_t) d\widetilde{\boldsymbol{\theta}}_t \\
&= \int_{-\infty}^{\infty} \begin{cases} \widetilde{\boldsymbol{\theta}}_t < a_t : 0 \\ \widetilde{\boldsymbol{\theta}}_t \geq a_t : 0.8 \end{cases} \\
&\qquad \otimes \begin{cases} \theta_{\tau_t} - \epsilon \leq \widetilde{\boldsymbol{\theta}}_t \leq \theta_{\tau_t} + \epsilon : \frac{1}{2\epsilon} \\ \widetilde{\boldsymbol{\theta}}_t < \boldsymbol{\theta}_{\tau_t} - \epsilon \vee \widetilde{\boldsymbol{\theta}}_t > \boldsymbol{\theta}_{\tau_t} + \epsilon : 0 \end{cases} d\widetilde{\boldsymbol{\theta}}_t \\
&= \begin{cases} \boldsymbol{\theta}_{\tau_t} \leq a_t - \epsilon : 0 \\ a_t - \epsilon < \boldsymbol{\theta}_{\tau_t} \leq a_t + \epsilon : 0.4(1 + \frac{\boldsymbol{\theta}_{\tau_t} - a_t}{\epsilon}) \\ \boldsymbol{\theta}_{\tau_t} > a_t + \epsilon : 0.8 \end{cases} \qquad (13)
\end{aligned}
$$

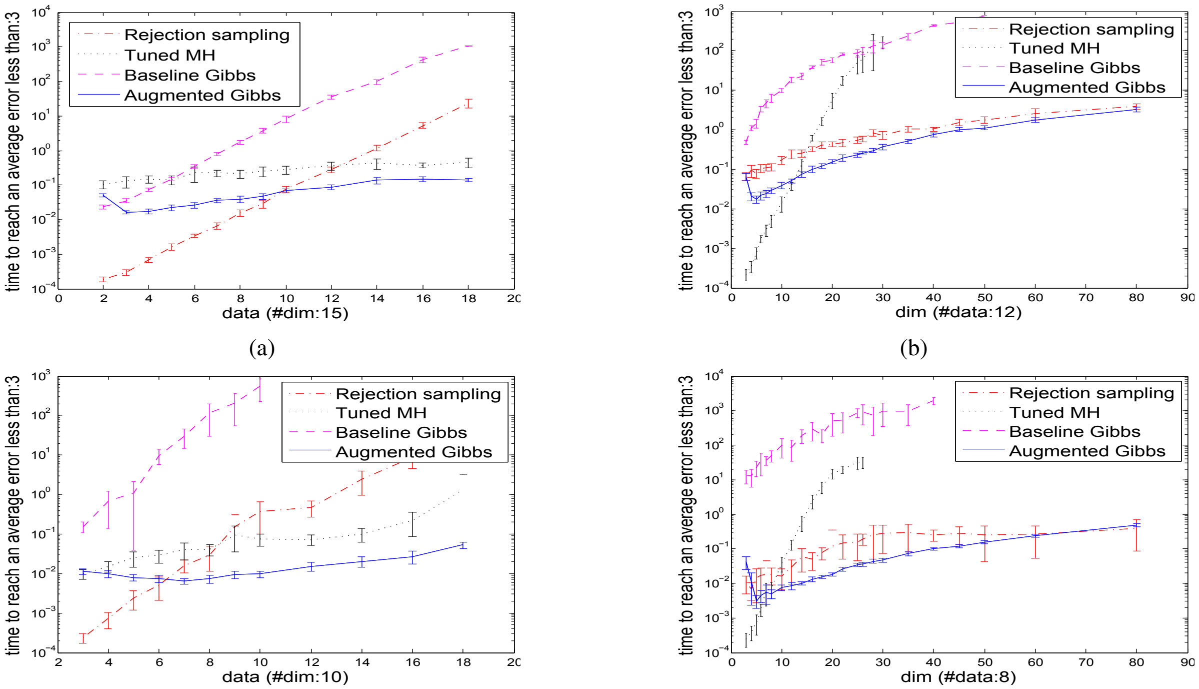
$\diamondsuit$

Models are configured as follows: In BPPL, $\eta = 0.4$ and *prior* is uniform in a hypercube. In MM, $L = 0$, $H = 20$, $\epsilon = 2.5$ and $\delta = 10$.

For each combination of the parameter space dimensionality N and the number of observed data n, we generate data points from each model and simulate the associated expected value of ground truth posterior distribution by running rejection sampling on a 4 core, 3.40GHz PC for 15 to 30 minutes. Subsequently, using each algorithm, particles are generated and based on them, average absolute error between samples and the ground truth, $||\mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}^*||_1$, is computed. The time till the absolute error reaches the threshold error 3 is recorded. For each algorithm 3 independent Markov chains are executed and the results are averaged. The whole process is repeated 15 times and the results are averaged and standard errors are computed.

We observe that in both models, the behavior of each algorithm has a particular pattern (Figure 5). The speed of rejection sampling and consequently its mixing time deteriorates rapidly as the number of observations increases.[2] The reason is that by observing new data, the posterior density tends to concentrate in smaller areas, leaving most of the space sparse and therefore hard to sample from by rejection sampling.

[2] For the same reason we were obliged to constraint the observations to less than 18 data points since beyond this, even after 30 minutes the number of samples generated to simulate the ground truth (which as mentioned were taken using rejection sampling) were not sufficient to guarantee a fair comparison between algorithms.



(a)

(b)

(c)

(d)

Figure 5: Performance of Rejection/Metropolis-Hastings, baseline and augmented Gibbs on (a) & (b) BPPL and (c) & (d) MM models against different configurations of the number of observed data and the dimensionality of the parameter space. In almost all cases, Augmented Gibbs takes orders of magnitude less time to achieve the same error as the other methods and this performance separation from competing algorithms increases in many cases with the amount of data and dimensionality.

It is known the efficiency of MH depends crucially on the scaling of the *proposal*. We carefully tuned MH to reach the acceptance rate 0.234 after the well known algorithm (Roberts, Gelman, and Gilks 1997). The experimental results show that MH is scalable in observations but its mixing time increases rapidly as the dimensionality increases. This results are rather surprising since we expected that as an MCMC, MH does not suffer from the curse of dimensionality as rejection sampling does. A reason may be that piecewise distributions can be non-smooth or broken which is far from the characteristics of the Gaussian *proposal density* used in MH.

Efficiency of the baseline Gibbs sampling in particular decreases as data points increase. The reason is that this leads to a blow up in the number of partitions in the posterior density. One the other hand, on both models, the augmented Gibbs is scalable in both data and dimension. Rather surprisingly, its efficiency even increases when dimensionality increases from 2 to 5. The reason may be that proportional to the total number of posterior partitions, in lower dimensions the neighbors of each partition are not so numerous. For instance, regardless of the number of observed data in the 2D BPPL, each partition is neighbored by only two partitions (see Figure 1) leading to a slow transfer between partitions. In higher dimensions however, this is not often the case.

Conclusion

In this work, we showed how to transform piecewise likelihoods in graphical models for Bayesian inference into equivalent mixture models and then provide a blocked Gibbs sampling approach for this transformed model that achieves an *exponential-to-linear* reduction in space and time compared to a conventional Gibbs sampler. Unlike rejection sampling and baseline Gibbs sampling, the time complexity of the proposed method does not grow exponentially with the amount of observed data. In both our test scenarios, its mixing time in higher dimensions is also better than Metropolis-Hastings. And in contrast to Metropolis-Hastings, it also does not require any tuning.

Future extensions of this work can also examine application of this work to non-Bayesian inference models (i.e., general piecewise graphical models). For example, some clustering models can be formalized as piecewise models with latent cluster assignments for each datum – the method proposed here allows linear-time Gibbs sampling in such models. To this end, this work opens up a variety of future possibilities for efficient asymptotically unbiased (Bayesian) inference in expressive piecewise graphical models that to date have proved intractable or inaccurate for existing (Markov Chain) Monte Carlo inference approaches.

References

[Das and Magdon-Ismail 2008] Das, S., and Magdon-Ismail, M. 2008. Adapting to a market shock: Optimal sequential market-making. In *NIPS*, 361–368.

[Guo and Sanner 2010] Guo, S., and Sanner, S. 2010. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *AI and Statistics (AISTATS)*.

[Keeney and Raiffa 1993] Keeney, R. L., and Raiffa, H. 1993. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

[Poon and Domingos 2006] Poon, H., and Domingos, P. 2006. Sound and efficient inference with probabilistic and deterministic dependencies. In *AAAI*, volume 6, 458–463.

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

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

[Shogren, List, and Hayes 2000] Shogren, J. F.; List, J. A.; and Hayes, D. J. 2000. Preference learning in consecutive experimental auctions. *American Journal of Agricultural Economics* 82(4):1016–1021.