Lifted Relational Variational Inferences

Jaesik Choi JAESIK@ILLINOIS.EDU
Eyal Amir EYAL@ILLINOIS.EDU
Computer Science Department,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801 USA

Abstract

We present a lifted inference algorithm for relational hybrid graphical models. Hybrid graphical models with continuous and discrete variables naturally represent many real-world applications in robotics, financial market predictions, and weather analysis. Inference with such large models is challenging because relational structures deteriorate rapidly with current inference procedures. The main contribution of this paper is a relational variational-inference lemma that enables factoring density functions into a mixture of independent identically distributed multi-valued Bernoulli trials. This lemma enables a relational factoring step that takes hybrid ground potentials and finds a close to optimal lifted relational model for the joint density. This step is then used for efficient inference without referring to ground random variables. The new method allows us to build various efficient inference algorithms. As an example, we provide a lifted Markov Chain Monte Carlo (MCMC) algorithm that requires fewer samples and generates each sample faster than possible before. We provide an error analysis of the variational method when applying to relational models. Our approach is applicable to general large relational models.

1. Introduction

Many real world systems can be described using continuous and discrete variables with relations among them. Such examples include measurements in environmental-sensors networks, localizations in robotics, and economic forecasting in finance. In large such systems, efficient and precise inference is essential. As an example from environmental science, an inference algorithm can predict a posterior of unobserved water levels and contamination levels at different locations, and making such inference precisely is critical to decision makers.

Relational Probabilistic Languages (RPLs) [14, 16, 10, 17, 7, 19, 13, 11, 2] describe probability distributions at a relational level with the purpose of capturing structure of larger models. These compact representations can facilitate the construction and learning of probabilistic models for large systems. A key challenge of inference procedures with RPLs is that they often result in density functions involving many random variables.

Recent advances (e.g. [7, 13, 11]) presented approaches to inference that (among others) group equivalent models into a histogram representation which includes an order of n^{k-1} entries (instead of performing k^{n-1} operations on traditional *ground* models)). Further, approximate lifted inference algorithms (e.g. [20, 1]) extended this approach and showed how to solve such inference problems with belief propagation and sampling (e.g. [21]).

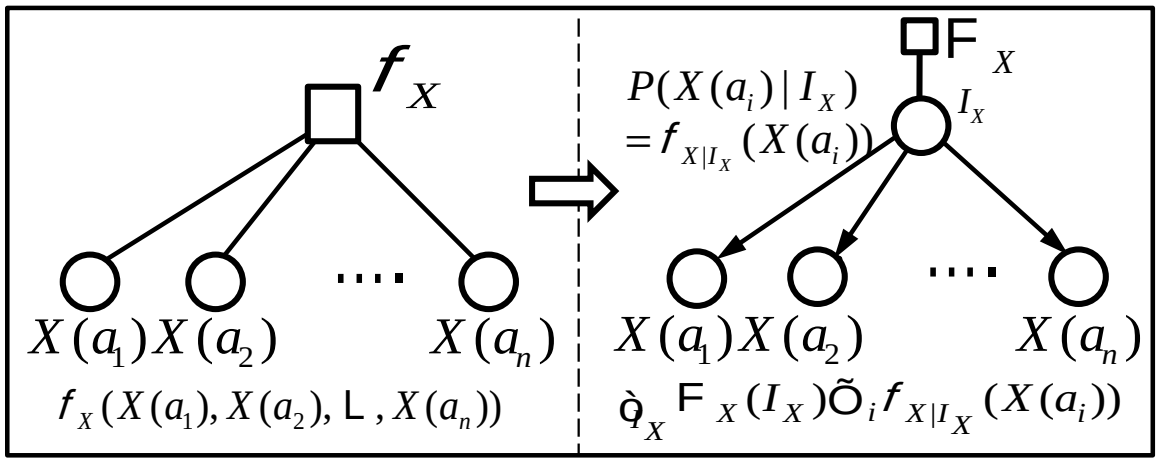


Figure 1: Illustration of a way to transform a potential with n exchangeable random variables $(X(a_1), \cdots, X(a_n))$ into a variational model with a latent variable I_X. The variational form (the right hand side) allows a compact representation with fewer parameters.

Unfortunately, these principles are not applicable to continuous or hybrid models, where k is large or infinite.

In this paper, we present an approach to relational lifted inference in Hybrid models. It applies a relational variational-inference lemma that we prove and that enables factoring density functions into a mixture of independent identically distributed multi-valued Bernoulli trials. This lemma enables a relational factoring step that takes hybrid ground potentials and finds a close to optimal lifted relational model for the joint density.

Our inference algorithm then can efficiently answer queries for large hybrid systems. First, it converts each potential in a relational model into a lifted variational form. The lifted variational model decouples ground random variables in a potential into a mixture of independent identically distributed Bernoulli trials. Then, lifted inference algorithms solve inference problem over the resulting models using this variational-approximation step. When density functions permit exact marginalization, an exact inference algorithm solves these problems. Otherwise, we use a lifted MCMC algorithm on those structures.

This paper is organized as follows. Section 2 provides the formal definition of Relational Hybrid Models (RHMs). Section 3 provides our basic result for relational variational steps. Section 4 explains how to calculate parameters of factored models, and gives the basic step for our followup algorithms. Section 5 shows how to apply this basic step in inference algorithms. Section 6 provides experimental results.

2. Relational Hybrid Models

A **Relational Hybrid Model (RHM)** is composed of a set F of factors. A **factor** f is a pair (A_f, ϕ_f) where A_f is a tuple of random variables and ϕ_f is a **potential function**, unnormalized probability density, from the range of A_f to the nonnegative real numbers. The domain of a random variable can be discrete or continuous, i.e. hybrid. Given a **valuation** v of random variables (**rv**s), the **potential** of f on v is $w_f(v) = \phi_f(A_f)$. The

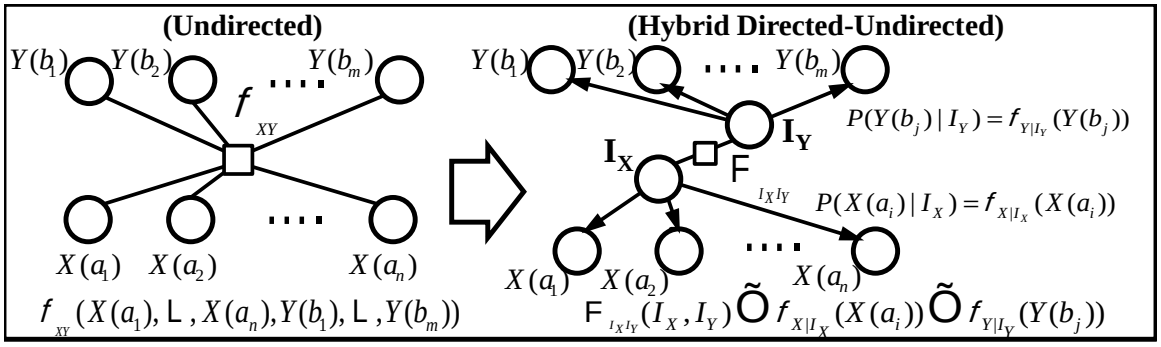


Figure 2: Illustration of a way to factor a potential with two relational atoms, $X(a_i)$ and $Y(b_j)$. Our variational method converts an undirected model (left) into a factor model (right). In the factored model, the distribution is dependent on the new latent variables, I_X and I_Y.

joint probability defined by a set F of factors on a valuation v of random variables is the normalization of $\prod_{f\in F} w_f(v)$.

An important property of the factor f is that its tuple A_f is a disjoint union of sets of **exchangeable random variables**[1] defined as follows:

Definition (**Exchangeable Random Variables).** A finite or infinite sequence $X(a_1),\cdots,X(a_n)$ of random variables are **exchangeable**, when for any finite permutation $\pi()$ of the indices the joint probability distribution of the permuted sequence $X(a_{\pi(1)}),\cdots,X(a_{\pi(n)})$ is the same as the joint probability distribution of the original sequence.

A **relational atom** refers a set of exchangeable random variables. For example, a potential with two relational atoms (or just atoms) $X()$ and $Y()$ can be represented as follows: $\phi(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$.

3. Relational Variational Lemma

A potential with a large number of random variables introduces inference difficulties of three kinds. First, it may require a large number of parameters to represent the probability density. Second, it is hard to learn the potential accurately unless a large number of training examples are given. Third, it requires a substantial amount of computations to marginalize out random variables participating in such potentials. To address these issues, we propose a model-factorization based variational method.

3.1 De Finetti-Hewitt-Savage's Theorem

Before introducing our variational method, we review de Finetti's theorem [6] which shows that any probability density function (pdf) of an infinite number of binary exchangeable

1. Note that, our representation is a general representation than existing relational models [15, 7, 19, 13, 21, 3, 11, 2] in terms of expressiveness. That is, not all potential withs exchangeable variables in RHMs can be represented by existing models either based on Parfactors [15, 7, 13, 3, 11, 2] or MNLs [19, 21].

random variables can be represented by a mixture of independent and identically distributed (**iid**) Bernoulli random variables (**RV**s) and the density over the RVs:

$$\lim_{n\to\infty} p(X(a_1),\cdots,X(a_n)) = \int_0^1 \theta^{t_n}(1-\theta)^{n-t_n}\cdot\Phi_X(\theta)\,\mathrm{d}\theta,$$

when $t_n = \sum_i X(a_i)$. This observation is extended to multi-valued RVs by Hewitt-Savage theorem.

$$\lim_{n\to\infty} p(X(a_1),\cdots,X(a_n)) = \int\prod_{i=1}^{n}\phi_{X|I_X}(X(a_i))\cdot\Phi_X(I_X)\,\mathrm{d}I_X \tag{1}$$

where I_X is a new latent variable which chooses a distribution $\phi_X(\cdot|I_X)$ for the iid multi-valued Bernoulli RVs.[2]. We also use $\phi_{X|I_X}(.)$ to refer to $\phi_X(.|I_X)$. Figure 1 illustrates an example of a variational form. The parameters of the potential (e.g. entries in the conditional density table (CDT)) could be substantially reduced by the factorization. Note that, the graphical model on the right hand side requires only a representation of $\phi_{X|I_X}$ and the density Φ_X over I_X.

This variational method is exact only if there is an infinite number of RVs. Thus, it is natural to analyze the error when we have a finite number of RVs. Before analyzing this error, we define a term $\overline{\mathbf{n}}$**-extendible**:

Definition ($\overline{\mathbf{n}}$**-extendible).** $p_n(X(a_1,\cdots,X(a_n))$, any pdf with n exchangeable RVs, is $\overline{\mathbf{n}}$**-extendible** when the following holds: (1) there is $p_{\overline{n}}(X(a_1),\cdots,X(a_n),X(a_{n+1}),\cdots,X(a_{\overline{n}}))$, a pdf with $\overline{n}$ exchangeable RVs ($n<\overline{n}$); and (2) p_n is the marginal distribution of $p_{\overline{n}}$ (i.e. eliminating $(\overline{n}-n)$ RVs).

Lemma 1 (**Diaconis and Freedman** [9])**.** *If $p_n(X(a_1),\cdots,X(a_n))$, any pdf with n exchangeable RVs, is $\overline{n}$-extendible, then the total variation distance $\|\cdot\|$ between p_n and the variational form in the Hewitt-Savage's theorem is bounded as follows: (i) when $X(a_i)$ are discrete RVs with a domain of cardinality c (e.g. c=2 for binary RVs), $\left\| p_n - \int\prod_{i=1}^{n}\phi_{X|I_X}(X(a_i))\cdot\Phi_X(I_X)\,\mathrm{d}I_X\right\| \le \frac{2cn}{\overline{n}}$; (ii) when $X(a_i)$ are continuous RVs, $\left\| p_n - \int\prod_{i=1}^{n}\phi_{X|I_X}(X(a_i))\cdot\Phi_X(I_X)\,\mathrm{d}I_X\right\| \le \frac{n(n-1)}{\overline{n}}$.*

The total variation distance is $\|p-q\| = sup_{A\in\mathcal{B}}(p(A)-q(A))$ when $\mathcal{B}$ is a class of Borel sets.

3.2 Factoring Potentials with Multiple Atoms

De Finetti's theorem and Heweitt-Savage's theorem of the previous section are applicable only to potentials with a single relational atom. In this section, we present our key new result that establishes variational methods for RHMs. Lemma 5 provides a result on potentials with an infinite number of objects. Lemma 6 provides an error bound on a single variational step, and Theorem 4 provides an error bound for our relational variational method in RHMs.

2. In general, the right hand side of Equation (1) is $\int\prod_{i=1}^{n}\phi_X(X(a_i)|I_X)\cdot\Phi_X(\,\mathrm{d}I_X)$. When the distribution Φ_X has a density, it is possible to replace $\Phi_X(\,\mathrm{d}I_X)$ with $\Phi(I_X)\,\mathrm{d}I_X$. Here, we only consider distributions of which density is defined.

Lemma 2 (**Existence of a variational factor).** *For $p_{n,m}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$, a potential with two relational atoms in a RHM, there are two new latent variables, I_X and I_Y, and a new potential of two variables p_{XY} such that the following holds,*

$$\lim_{n,m\to\infty} p_{n,m}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m)) = \int \Phi_{XY}(I_X,I_Y)\prod_{i=1}^{n}\phi_{X|I_X}(X(a_i))\prod_{j=1}^{m}\phi_{Y|I_Y}(Y(b_j))\,\mathrm{d}I_X\,\mathrm{d}I_Y.$$

We extend the previous framework and define the term $(\overline{\mathbf{n}},\overline{\mathbf{m}})$**-extendible**. It then allows us to derive an error analysis for $p_{n,m}$[3]:

Definition ($(\overline{\mathbf{n}},\overline{\mathbf{m}})$**-extendible).** $p_{n,m}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$, any pdf with n exchangeable RVs of $X()$ and m exchangeable RVs of $Y()$ is $(\overline{\mathbf{n}},\overline{\mathbf{m}})$**-extendible** when it holds followings: (1) there is $p_{\overline{n},\overline{m}}(X(a_1),\cdots,X(a_{\overline{n}}),Y(b_1),\cdots,Y(b_{\overline{m}}))$, a pdf with $\overline{n}$ exchangeable RVs of $X(\cdot)$ and $\overline{m}$ exchangeable RVs of $Y(\cdot)$ ($n<\overline{n}, m<\overline{m}$); and (2) $p_{n,m}$ is the marginal distribution of $p_{\overline{n},\overline{m}}$ (i.e. eliminating $(\overline{n}-n)$ RVs of $X()$ and $(\overline{m}-m)$ RVs of $Y()$).

Lemma 3 (**Error of a variational factor).** *If $p_{n,m}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$, any pdf with two relational atoms, in a RHM is $(\overline{n},\overline{m})$-extendible, the total variation distance of $p_{n,m}$ and the $p_{iid(n,m)}$ (the variational form in Lemma 5) is bounded as follows: (i) when RVs of $X()$ and $Y()$ are discrete with domains of cardinality c_x and c_y respectively, $\left\|p_{n,m}-p_{iid(n,m)}\right\| \le \frac{2c_x n}{\overline{n}}+\frac{2c_y m}{\overline{m}}$; (ii) when $X(\cdot)$ are discrete RVs with a domain of cardinality c_x and $Y(\cdot)$ are continuous RVs, $\left\|p_{n,m}-p_{iid(n,m)}\right\| \le \frac{2c_x n}{\overline{n}}+\frac{m(m-1)}{\overline{m}}$; (iii) when $X()$ and $Y()$ are continuous RVs, $\left\|p_{n,m}-p_{iid(n,m)}\right\| \le \frac{n(n-1)}{\overline{n}}+\frac{m(m-1)}{\overline{m}}$.*

For pdfs with more than two relational atoms (e.g. $p(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m),Z(c_1),\cdots,Z(c_u))$, it is natural to extend Lemma 6 as follows: The total variation distance is bounded by the sum of the variation distances for all relational atoms in the potential.

Theorem 4 (***Variational error of a RHM*).** *For each factor f_i in a RHM F, its potential p_i is a pdf (or normalized), and the total variation distance between p_{f_i} and its variational form $p_{iid(i)}$ is bounded by ϵ_i, then the total variation distance between the joint distribution of G and its variational form is bounded by $\frac{1}{z}\sum_i \epsilon_i$ when z is the normalizing constant of $\prod_i p_{f_i}$.*

We provide proofs of Lemma 5, Lemma 1, and Theorem 4 in Appendix A.1, A.2 and A.3.

4. How to Find a Variational RHM?

The previous section concerned the existence of a relational form that represents or approximates our original distribution well. In this section[4] we address the question of how to convert potentials (e.g. p_n) into a variational lifted relational form.

When a given potential is ∞-extendible, it is possible to derive the cdf $F_X(I_X)$ on I_X analytically as follows [8]: $F_X(I_X)=\lim_{n\to\infty}\frac{1}{n}|\{i|X(a_i)\le I_X\}|$. For example, relational Models

3. We use $p_{n,m}$ to refer the original (unfactored) potential ϕ_{XY}, and $p_{iid(n,m)}$ to refer the factored model. $p_{n,m}$ and $p_{iid(n,m)}$ are the pdfs of the potential forms $\phi_{n,m}$ and $\phi_{iid(n,m)}$, respectively. The difference between potential and pdf is that a pdf is integrated into 1 while a potential is not.
4. A variational RHM is a RHM in which all potentials are converted into the variational forms

with continuous RVs (e.g. pairwise Gaussian [3] and Gaussian processes [4, 23]) allow such analytical derivations [12][5].

Unfortunately, it is not possible to use such derivations in general because some potentials are not ∞-extendible, so $\lim_{n\to\infty} \frac{1}{n} |\{i|X(a_i) \leq I_X\}|$ is not defined.

In the following we focus on providing solutions for non-trivial problems which have no analytical solution. Here, we provide solutions for discrete models first, then for continuous models.

4.1 Lifting Discrete Variables

When an RHM includes potentials with discrete RVs, we need to find a density function $\Phi_{X(I_X)}$ over the iid Bernoulli RVs of parameters I_X. We start with binary RVs. To solve the problem, we formulate Equation (1) as follows:

$$\arg\max_{\Phi_X(I_X)} \left\| \phi_{\mathbf{h}}(h_X) - \int \Phi_X(I_X) \cdot f_{\mathrm{B}}(h_X; n, I_X)\, \mathrm{d}I_X \right\| \approx \arg\max_{\langle (w_1, i_X^1), \cdots, (w_k, i_X^k) \rangle} \left\| \phi_{\mathbf{h}}(h_X) - \sum_{l=1}^{k} w_l \cdot f_{\mathrm{B}}(h_X; n, i_X^l) \right\|, \tag{2}$$

where $\sum_{l=1}^{k} w_l = 1$. $\phi_{\mathbf{h}}(h_X)$ in Equation (2) is the value-histogram representation introduced in previous lifted-inference methods [7, 13]. Thus, given values of n RVs, $X(a_1), \cdots, X(a_n)$, the value histogram is a vector h_X with $h_{X_v} = |\{i : X(a_i) = v\}|$ for each v in RVs' range. When h_X is the value histogram of X, $\phi_n(X(a_1), \cdots, X(a_n)) = \phi_{\mathbf{h}}(h_X)$. $f_B(h_X; n, I_X) = \binom{n}{h_{X_1}} I_X^{h_{X_1}} (1 - I_X)^{n - h_{X_1}}$ is the pdf of the binomial distribution.

The approximation in Equation (2) is due to our incremental iterative algorithm, choosing an empirical k through iterations.[6]

For binary RVs, $\phi_{X|I_X}$ is the Bernoulli (distribution) with I_X^l as a parameter (i.e. $P(X(a_j){=}1) = i_X^l$). Thus, i_X^l in Equation (2) is a parameter of the binomial distribution, and w_l is the density on the Bernoulli distributions. That is, the problem is to find a mixture of binomial distributions.

For multi-valued variables, $\phi_{X|I_X}$ is the multi-valued Bernoulli, i.e. Categorical distribution. The problem is to find a mixture of multinomial distributions f_{M}:

$$\arg\max_{\langle (w_1, i_X^1), \cdots, (w_k, i_X^k) \rangle} \left\| \phi_{\mathbf{h}}(h_X) - \sum_{l=1}^{k} w_l \cdot f_{\mathrm{M}}(h_X; n, i_X^l) \right\| \tag{3}$$

For potentials with two or more relational atoms, it can be formulated as follows:

$$\arg\max_{\langle (w_1, i_X^1, i_Y^1), \cdots, (w_k, i_X^k, i_Y^k) \rangle} \left\| \phi_{\mathbf{h}}(h_X, h_Y) - \sum_{l=1}^{k} w_l \cdot f(h_X; n, i_X^l) \cdot f(h_Y; m, i_Y^l) \right\|, \tag{4}$$

where f is either the binomial or the multinomial which depends on the range of RVs.

We learn the parameters (e.g. (w, i_X) in Equation (2)) from the original potential ϕ using an EM algorithm that solves Equations (2), (3) and (4).[7] Normally, such EM algorithms assume that k is known or given. Because the assumption does not hold for this case, we increase k with an incremental EM algorithm until the error converges. The achieved error

5. For Gaussian processes with an infinite number of exchangeable RVs, the mean of RVs follows a Gaussian distribution. Given a mean, each RV also follows a Gaussian distribution.
6. Notice that this is applicable even when there is no analytical derivation for the model
7. EM algorithms are used to learn parameters for mixture models, e.g. [22, 5, 18].

with k components is the empirical error of the theoretical one in Lemma 6. It is well known that EM algorithms derive a close-to-optimal mixture model when components in the true density are well separated (e.g. $|\mu_i - \mu_j| > \sigma_i^2 + \sigma_j^2$ for Gaussian mixtures) [22].

4.2 Lifting Continuous Variables

For a potential ϕ with RVs in a continuous domain, finding the variational lifted relational form requires additional considerations for non-parametric densities. The variational form for continuous domains is possibly a mixture of non-parametric densities. Here, we generate samples from the input potential ϕ, then learn a mixture of non-parametric densities.

Equation (1) for discrete potentials is adapted to continuous and hybrid potentials as follows:

$$\arg\max_{\Phi_X(I_X)} \left\| \phi_n(X()) - \int \Phi_X(I_X) \cdot \prod_{j=1}^{n} \hat{f}_{I_X}(X(a_j))\, \mathrm{d}I_X \right\| \approx \arg\max_{\langle (w_1, \hat{f}_{i^1_X}), \cdots, (w_k, \hat{f}_{i^k_X}) \rangle} \left\| \phi_n(X()) - \sum_{l=1}^{k} w_l \cdot \prod_{j=1}^{n} \hat{f}_{i^l_X}(X(a_j)) \right\|, \tag{5}$$

where $\phi_n(X()) = \phi_n(X(a_1), \cdots, X(a_n))$ and $\hat{f}_{I_X}$ represents a non-parametric distribution. To solve this equation, we generate N samples $V^1, \cdots, V^N$ from the input potential ϕ where $V^j = (v_1^j, \cdots, v_n^j)$, i.e. values of n RVs. Then, the problem is formulated as the maximum likelihood estimation (MLE) problem: $\arg\max_{\langle (w_1, \hat{f}_{i^1_X}), \cdots, (w_k, \hat{f}_{i^k_X}) \rangle} \sum_{t=1}^{N} ln\left(\sum_{l=1}^{k} w_l \cdot \prod_{j=1}^{n} \hat{f}_{i^l_X}(v_j^t) \right)$.

We denote by $\hat{f}_{i^j_X}$ the kernel density estimator: $\hat{f}_{i^l_X} = \frac{1}{n_l h} \sum_{s=1}^{n_l} K\left(\frac{x - v_s}{h}\right)$ when $(v_1, \cdots, v_{n_l})$ are data points that underlie the density and h is the parameter. Here, we use a Gaussian Kernel, $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. It is interesting to note that the kernel density estimator is analogous to the value histogram of discrete RVs in a sense that frequently observed regions (or bins) have higher probability. In this way, the intuition of the value histogram helps us generalize the method for continuous RVs.

For potentials with two or more relational atoms, the approach can be formulated as follows: $\arg\max_{\langle (w_1, \hat{f}_{i^1_X}, \hat{f}_{i^1_Y}), \cdots, (w_k, \hat{f}_{i^k_X}, \hat{f}_{i^k_Y}) \rangle} \sum_{t=1}^{N} ln\left(\sum_{l=1}^{k} w_l \cdot \prod_{j=1}^{n} \hat{f}_{i^l_X}(v^t_{X_j}) \cdot \prod_{j'=1}^{m} \hat{f}_{i^l_Y}(v^t_{Y_{j'}}) \right)$, where $v^t_{X_j}$ is the value of j-th RV of X in the t-th sample, and $v^t_{Y_{j'}}$ is the value of j'-th RV of Y.

This MLE problem can also be solved by an EM algorithm. There, one of N samples will be used to build one of k densities in a maximization (M) step, and the likelihood of each sample from k densities is calculated in an expectation (E) step. With discrete RVs, k is determined in an incremental way until the variation error converges.

5. Lifted Inference with Variational RHM

Our previous two sections presented methods and error bounds for lifting and factoring complex hybrid relational models. Those results provide variational RHMs in which all potentials are in variational relational lifted forms.

In this section we build on those results and present two algorithms that apply the variational steps above to speed up relational inference algorithms. Sections 5.1 and 5.2 present Lifted Hybrid Variable Elimination on those models. Section 5.3 addresses the case when the resulting model is still too complex for exact inference, and present an MCMC sampling algorithm that samples at a lifted level without grounding unnecessarily.

5.1 Inference with Discrete Variables

A variable elimination (VE) is an inference procedure with following steps: (i) choosing an atom; (ii) finding all potential including the atom; (iii) making a produce of the found potentials; (iv) marginalizing the atom; and (v) repeating the steps until only output atoms remain. We demonstrate the key step in our Lifted Variational VE, Step (iv), with an example: a potential ϕ with two atoms $X()$ and $Y()$ and ϕ' with a atom $Y()$. Suppose that the potentials are converted into a variational form as Equation (4). Then the marginal probability of I_X can be derived by marginalizing the atom $Y()$ out: $\sum_{h_y} \phi_{\mathbf{h}}(h_x, h_y) \cdot \phi'_{\mathbf{h}}(h_y)$

$$\approx \sum_{h_y} \sum_{l=1}^{k} w_l f(h_x; n, i^l_X) f_{\mathrm{B}}(h_y; m, i^l_Y) \sum_{l'=1}^{k'} w_{l'} f_{\mathrm{B}}(h_y; m, i^{l'}_Y) = \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l w_{l'} \left(\sum_{h_y} f_{\mathrm{B}}(h_y; m, i^l_Y) f_{\mathrm{B}}(h_y; m, i^{l'}_Y) \right) f(h_x; n, i^l_X)$$

$$\approx \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l w_{l'} \left(\int f_{\mathrm{N}}(h_y; \mu_{(m,i^l_Y)}, \sigma^2_{(m,i^l_Y)}) f_{\mathrm{N}}(h_y; \mu_{(m,i^{l'}_Y)}, \sigma^2_{(m,i^{l'}_Y)}) \, \mathrm{d}h_y \right) f(h_x; n, i^l_X) = \sum_{l=1}^{k} w_l^{\overline{Y}} \cdot f(h_x; n, i^l_X) \quad (6)$$

when $\sum_{l=1}^{k} w_l^{\overline{Y}} = 1$, where $f_{\mathrm{N}}(h_y; \mu_{(n,p)}, \sigma^2_{(n,p)})$ is a pdf of the Normal distribution with a mean $\mu_{(n,p)}(= n \cdot p)$ and a variance $\sigma^2_{(n,p)}(= n \cdot p \cdot (1-p))$, and $z_{l,l'}$ is the inverse of the normalizing constant calculated from the product of two Normal pdfs. Equation (6) is the Normal approximation to Binomial for a large m. When m is small, we can keep the value histogram representation. Then, the marginal density is still represented as a mixture of iid Bernoulli RVs. Note that, other procedures involving more relational atoms hold the property, although the representation may include more components.

Now, we will show that the product of variational forms in Step (iii) can also be represented as a variational from. Suppose that we have two probabilities for I_X of binary RVs, one from $\phi_{\mathbf{h}}(h_x, h_y)$ after marginalizing $Y()$ out, and another from $\phi'_{\mathbf{h}}(h_x, h_z)$ after marginalizing $Z()$ out, i.e. $\sum_{l=1}^{k} w_l^{\overline{Y}} \cdot f_{\mathrm{B}}(h_x; n, i^l_X)$ and $\sum_{l'=1}^{k'} w_{l'}^{\overline{Z}} \cdot f'_{\mathrm{B}}(h_x; n, i^{l'}_X)$. Note that, it can be $k \neq k'$ and $f_{\mathrm{B}}(h_x; n, i^l_X) \neq f'_{\mathrm{B}}(h_x; n, i^l_X)$ because the parameters are extracted from different potentials, $\phi_{\mathbf{h}}(h_x, h_y)$ and $\phi'_{\mathbf{h}}(h_x, h_z)$. Then, the product of two potentials $\phi_{\mathbf{h}}(h_x)$ and $\phi'_{\mathbf{h}}(h_x)$ is as follows:

$$\sum_{l=1}^{k} w_l^{\overline{Y}} \cdot f_{\mathrm{B}}(h_x; n, i^l_X) \cdot \sum_{l'=1}^{k'} w_{l'}^{\overline{Z}} \cdot f'_{\mathrm{B}}(h_x; n, i^{l'}_X) = \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \cdot \sum_{h_x} f_{\mathrm{B}}(h_x; n, i^l_X) \cdot f'_{\mathrm{B}}(h_x; n, i^{l'}_X)$$

$$\approx \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \int f_{\mathrm{N}}(h_x; \mu_{(n,i^l_X)}, \sigma^2_{(n,i^l_X)}) \cdot f'_{\mathrm{N}}(h_x; \mu_{(n,i^{l'}_X)}, \sigma^2_{(n,i^{l'}_X)}) \, \mathrm{d}h_x = \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \cdot z_{l,l'} f_{\mathrm{N}}(h_x; \mu_{new}, \sigma^2_{new}), \quad (7)$$

$z_{l,l'}$ is the inverse of the normalization constant of the product of two Normal pdfs. This has $|k \cdot k'|$ Normal components, but there are several way to reduce some of them, which we omit here for lack of space.

5.2 Inference with Continuous Variables

Similar to the discrete cases, we demonstrate Lifted Variational VE for continuous variables with an example. Assume two potentials ϕ with two atoms $X()$ and $Y()$, and ϕ' with a atom $Y()$. Two potentials are represented by a variational form such as in Equation (5). Then the marginal probability of $X()$ can be derived by integrating the atom $Y()$ out as follows:

$\int \cdots \int \phi(X(), Y()) \cdot \phi'_{\mathbf{h}}(Y()) \, \mathrm{d}Y(b_1) \cdots \mathrm{d}Y(b_m)$.

$$\int \cdots \int \left(\sum_{l=1}^{k} w_l \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)) \prod_{j'=1}^{m} \hat{f}_{i_Y^l}(Y(b_{j'})) \right) \left(\sum_{l'=1}^{k'} w_{l'} \prod_{j''=1}^{m} \hat{f}_{i_Y^{l'}}(Y(b_{j''})) \right) \mathrm{d}Y(b_1) \cdots \mathrm{d}Y(b_m)$$

$$= \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l w_{l'} \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)) \cdot \prod_{j'=1}^{m} \left(\int \hat{f}_{i_Y^l}(Y(b_{j'})) \cdot \hat{f}_{i_Y^{l'}}(Y(b_{j'})) \, \mathrm{d}Y(b_{j'}) \right)$$

$$= \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l w_{l'} \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)) \prod_{j'=1}^{m} z_{l,l'} = \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l w_{l'} {z_{l,l'}}^m \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)), \tag{8}$$

$z_{l,l'}$ is the inverse of the normalization constant of the product of two mixture of Normals, $\hat{f}_{i_Y^l}(Y(b_{j'}))$ and $\hat{f}_{i_Y^e}(Y(b_{j'}))$.

Finally, notice that for RVs of continuous domains the product of two variational forms has a variational form: $\left(\sum_{l=1}^{k} w_l^{\overline{Y}} \cdot \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)) \right) \cdot \left(\sum_{l'=1}^{k'} w_{l'}^{\overline{Z}} \cdot \prod_{j=1}^{n} \hat{f'}_{i_X^{l'}}(X(a_j)) \right)$

$$= \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \prod_{j=1}^{n} \hat{f}_{i_X^l}(X(a_j)) \cdot \hat{f'}_{i_X^{l'}}(X(a_j)) = \sum_{l=1}^{k} \sum_{l'=1}^{k'} w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \cdot {z_{l,l'}}^n \cdot \prod_{j=1}^{n} \hat{f}^{new}_{i_X^{l,l'}}(X(a_j)). \tag{9}$$

5.3 Lifted Variational Markov chain Monte Carlo (MCMC)

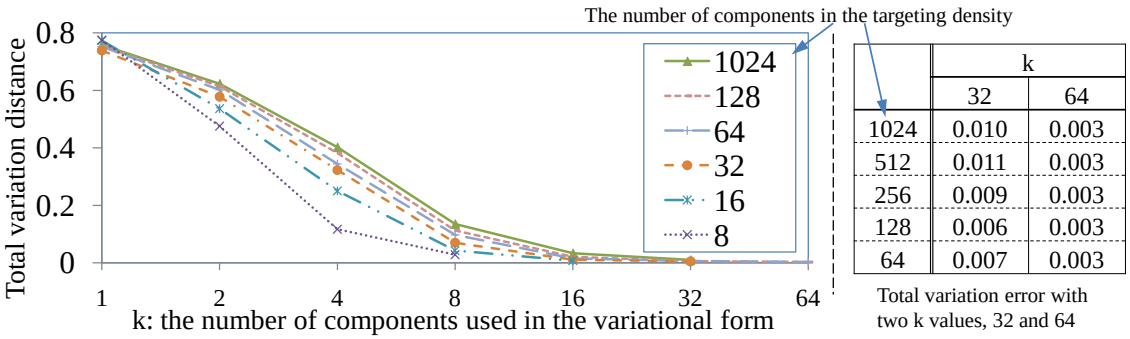
When lifted relational variational relational models are still too complex for Lifted Variational VE, we can apply MCMC on the results of our lifting operatios above. In general, MCMC sampling is composed of five steps: (i) choosing a RV to sample; (ii) calculating the conditional probability of each potential using assignments of neighboring RVs; (iii) building a probability a density with the product of the conditional probabilities; (iv) choosing an assignment from the density; and (v) repeating until some conditions are met.

Here, the main steps are Steps (ii), (iii), and (iv). Step (ii) is a subset of the marginalization procedure in Equations (6) and (8). Step (iii) can be derived in a straightforward manner from Equations (7) and (9).

Thus, we focus our attention on Step (iv). Recall that we choose a component according to the results in Equations (7) and (9). Essentially, we choose one component for $X()$ proportional to $w_l^{\overline{Y}} \cdot w_{l'}^{\overline{Z}} \cdot z_{l,l'}$ out of $|k| \cdot |k'|$ Normal pdfs. With Lifted MCMC (compared with ground MCMC) we also need to choose a tuple of indices for all potentials which include $X()$. For example, when we choose the (l, l')-th component, we assign a tuple of indices $(i_X^l, i_X^{l'})$ for I_X. Then, the first index i_X^l will be used to calculate the conditional probability of $\phi_{\mathbf{h}}(h_y)$ in $\phi_{\mathbf{h}}(h_x, h_y)$, and $i_X^{l'}$ will be used in $\phi_{\mathbf{h}}(h_x, h_z)$.

6. Experimental Results

We provide experimental results about the number of components in the lifted relational variational form, and the computational efficiency of the lifted inference. First, we address the question: to what degree can we reduce the number of components in the variational form. To examine this, we build a simulation with a single atom of 100 RVs. We randomly choose 30 mixtures of Binomials (various numbers from 8 to 1024) per parameters. Figure 3 shows the average variational distance of the target density and our variational

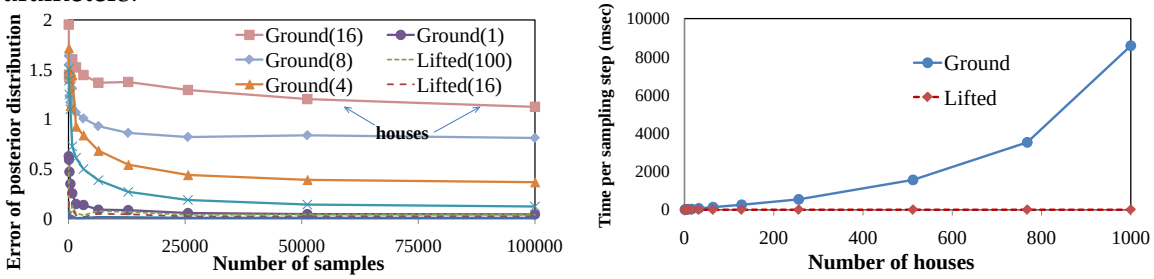


	k	
	32	64
1024	0.010	0.003
512	0.011	0.003
256	0.009	0.003
128	0.006	0.003
64	0.007	0.003

Total variation error with two k values, 32 and 64

Figure 3: The variation distance of our lifted relational variational model with k components. Even if a target density has a larger number of components, we obtain a close approximate density with a reasonably small k.

model derived from our EM algorithm. It shows that with a significant fewer number of components (e.g. 32) the variation distance becomes reasonable small (≤ 0.01). When we increase the number RVs (e.g. 200 and 1000), the results are consistent with the plot. Thus, it shows that it is a reasonable idea to use the incremental iterative algorithm to learn the parameters.



(a) The accuracy of samplings

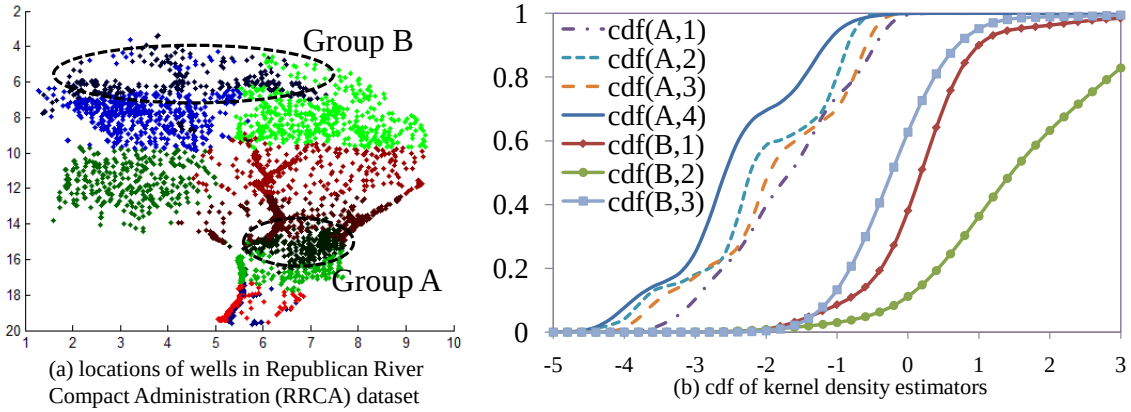
(b) The sampling time

Figure 4: Figure (a) compares the accuracy of posterior distributions of our lifted MCMC and the ground MCMC with various numbers of houses. '()' indicates the number of houses (e.g. 'Ground(16) is a ground MCMC with 16 houses). Figure (b) shows the average sampling time per each time step with various number of houses.

Second, we examine the computational improvement of our algorithms compared with ground algorithms. We compare the accuracy and the efficiency of our lifted MCMC algorithm with a traditional (ground) MCMC algorithm on a linear Gaussian model. The model is composed of two relational atoms $Job()$ and $HPC()$ (or $HousePriceChange()$). The ϕ_{Job} is the Bernoulli distribution with a parameter p_{Job}. The ϕ_{HPC} is a mixture of two Gaussians: $w_{DN}N(-0.3, \sigma^2_{DN})+w_{UP}N(0.1, \sigma^2_{UP})$. Then, the parameters of two atoms are related by the following linear Gaussian: $\Phi(\phi_{Job}, \phi_{HPC})=N(p_{job}-w_{DN}, \sigma^2_{JH})$. Figure 4a shows the accuracy of the two algorithm given the same number of samples. That is, it measure the ratio of error to estimate a probability density of an event x, $|p_{true}(x)-p_{MCMC}(x)|/p_{true}(x)$. It shows that the ground MCMC suffer from the curse of dimensionality, when the search space is lager. Meanwhile, the lifted MCMC converges to the true density quickly. Figure

4b represent the sampling time per step with different number of RVS (e.g. the number of houses).

Finally, we find an exemplar model in Republican River Compact Administration (RRCA) dataset.[8] RRCA Ground Water Model (RRCA Model) is to determine the amount, location, and timing of streamflow depletions to the Republican River caused by various effects such as well pumping. However, the state-of-the-art RRCA model is not always accurate. Thus, it is required to compensate the error (or residual) of estimation in each well. In a preliminary experiment, we cluster the locations of wells into 10 groups which shows a similar (approximately exchangeable) residuals pattern. Then, we select two of regions, groups (or atoms) A and B. Figure 5 shows identified cdfs for each groups. As the discrete case, with only small number of components (4 for A and 3 for B), we can represent the mixtures of cdfs. That is, there is no substantial improvement of the likelihood when we increase the number of mixtures.



(a) locations of wells in Republican River Compact Administration (RRCA) dataset

(b) cdf of kernel density estimators

Figure 5: Figure (a) shows the locations of clustered wells in the regions. Figure b) represented Gaussian kernel density estimators learned by our EM algorithm.

7. Conclusion

We propose an efficient lifted inference algorithm for RHMs with discrete variables and continuous variables. With a variational method, we reduce the time and space complexity of handling potentials with a large number of RVs. It is the first variational lifted inference algorithm which is generally applicable to various types of potentials in hybrid domains. Thus, it is a scalable algorithm which can be used for intractable hybrid graphical models with a large number of RVs.

8. Acknowledgement

This work is supported by NSF EAR 09-43627 EA.

8. Republican River Compact Administration, http://www.republicanrivercompact.org/.

Appendix A. Additional Proofs

A.1 Existence of a variational factor

Lemma 5 (Existence of a variational factor)**.** *For a potential with two relational atoms in a RHM, $\phi_{XY}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$, there are two new latent variables, I_X and I_Y, and a new potential of two variables Φ_{XY} such that the following holds,*

$$\lim_{n,m\to\infty}\phi_{XY}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))=\int\Phi_{XY}(I_X,I_Y)\prod_i\phi_{X|I_X}(X(a_i))\prod_j\phi_{Y|I_Y}(Y(b_j))\,dI_X\,dI_Y.$$

Proof. Given the potential ϕ_{XY}, suppose that the value of $X(a_1),\cdots,X(a_n)$ are assigned with constants $(c_1,\cdots,c_n)$. Then, from the Hewitt-Savage's theorem, it can be factored as follows.

$$\begin{aligned}&\lim_{m\to\infty}\phi_{XY}(c_1,\cdots,c_n,Y(b_1),\cdots,Y(b_m))\\ =\;&\int\Phi_Y(I_Y)\prod_j\phi_{Y|I_Y}(Y(b_j))\,dI_Y.\end{aligned}\tag{10}$$

Note that, $\Phi_Y(I_Y)$ is a pdf over I_Y. That is, $\Phi_Y(c_Y)$ is a constant density for an assignment, $I_Y=c_Y$. By an assignment for n RVs $(X(a_i)\cdots,X(a_n))$, the constant density $\Phi_Y(c_Y)$ can be represented as a function of the n RVs for a further factoring.

$$\begin{aligned}\lim_{n\to\infty}\Phi_Y(c_Y)&=\lim_{n\to\infty}\phi_{c_Y}(X(a_1),\cdots,X(a_n))\\&=\int\Phi_{Xc_Y}(I_X)\prod_i\phi_{X|I_X}(X(a_i))\,dI_X\end{aligned}$$

To represent general cases, it is enough to allow that the c_Y in $\Phi_{Xc_Y}(I_X)$ is parameterized by I_Y ($\Phi_{XY}(I_X,I_Y)$) as follows.

$$\lim_{n\to\infty}\Phi_Y(Y)=\int\Phi_{XY}(I_X,I_Y)\prod_i\phi_{X|I_X}(X(a_i))\,dI_X.\tag{11}$$

When we substitute $\Phi_Y(Y)$ in Equation (10) with Equation (11), the following result is derived.

$$\int\Phi_{XY}(I_X,I_Y)\prod_i\phi_{X|I_X}(X(a_i))\prod_j\phi_{Y|I_Y}(Y(b_j))\,dI_XI_Y.\quad\square$$

A.2 Error of a variational factor

Lemma 6 (Error of a variational factor). *If $p_{n,m}(X(a_1),\cdots,X(a_n),Y(b_1),\cdots,Y(b_m))$, any pdf with two relational atoms, in a RHM is $(\overline{n},\overline{m})$-extendible, the total variation distance of $p_{n,m}$ and the $p_{iid(n,m)}$ (the variational form in Lemma 5) is bounded as follows: (i) when RVs of $X()$ and $Y()$ are discrete with domains of cardinality c_x and c_y respectively, $\left\|p_{n,m}-p_{iid(n,m)}\right\|\le\frac{2c_xn}{\overline{n}}+\frac{2c_ym}{\overline{m}}$; (ii) when $X(\cdot)$ are discrete RVs with a domain of cardinality c_x and $Y(\cdot)$ are continuous RVs, $\left\|p_{n,m}-p_{iid(n,m)}\right\|\le\frac{2c_xn}{\overline{n}}+\frac{m(m-1)}{\overline{m}}$; (iii) when $X()$ and $Y()$ are continuous RVs, $\left\|p_{n,m}-p_{iid(n,m)}\right\|\le\frac{n(n-1)}{\overline{n}}+\frac{m(m-1)}{\overline{m}}$.*

Proof. We need to review the proof for a single atom in [9]. It shows that a $\bar{n}$-extendible pdf, p_n, can be represented by a mixture of extreme pdf (e.g. $p_n = \sum_e w_e p_e$). Here, an extreme pdf p_e is a distribution of n draws made at random without replacement from an urn, U, which contains $\bar{n}$ balls marked by one of c colors. Let e a unique marking in U. The variation distance of each extreme point p_e and its variational form $\prod_i \phi_{X|e}(X_i)$ is bounded $\leq \frac{2cn}{\bar{n}}$ for discrete RVs. .

For a distribution with the multiple atoms, each extreme point corresponds to the joint distribution of n draws from one urn, U_X of $\bar{n}$ balls, and m draws from another urn, U_Y of $\bar{m}$ balls, respectively. The draws can be done independently for each urn. Thus, an extreme pdf (e.g. p_{e_x,e_y}) can be represented as the product of independent extreme pdfs (e.g. $p_{e_x} \cdot p_{e_y}$). The variation distances of variational forms of p_{e_x} and p_{e_y} are respectively bounded. WLOG, we can represent the errors with ϵ_x and ϵ_y,

$$\left\| p_{e_x} - \prod_i \phi_{X|e_x}(X_i) \right\| \leq \epsilon_x, \left\| p_{e_y} - \prod_j \phi_{Y|e_y}(Y_j) \right\| \leq \epsilon_y.$$

Then,

$$\left\| p_{e_x} \cdot p_{e_y} - \prod_i \phi_{X|e_x}(X_i) \cdot \prod_j \phi_{Y|e_y}(Y_j) \right\| \leq \epsilon_x + \epsilon_y.$$

Thus, total variation distance between two densities, $p_{n,m}$ and $p_{iid(n,m)}$ is bounded by the product of the sum of two error bounds and the normalization constant ($\frac{1}{z}$). Note that, the product of two pdfs may not be a pdf without a normalization constant. □

A.3 Variational error of a RHM

Theorem 7 (***Variational error of a RHM***)**.** *For each factor f_i in a RHM F, its potential p_i is a pdf (or normalized), and the total variation distance between p_{f_i} and its variational form $p_{iid(i)}$ is bounded by ϵ_i, then the total variation distance between the joint distribution of F and its variational form is bounded by $\frac{1}{z}\sum_i \epsilon_i$ when z is the normalizing constant of $\prod_i p_{f_i}$.*

Proof. WLOG, we refer relational atoms in F as $X_1, \cdots, X_N$ when N is the number of relational atoms. Here, we shorten the j_{th} rv of the i_{th} atom $X_i(a_j)$ into X_i^j. n_i refers to the number of RVs in the i_{th} atom (i.e. $X_i^1, \cdots, X_i^{n_i}$).

The joint distribution of all relational atoms can be written as the product of pdfs in factors. Thus, it is possible to build an aggregated pdf p with all relational atoms as arguments (unfactored form).

$$p(X_1^1, \cdots, X_1^{n_1}, X_2^1, \cdots, X_2^{n_2}, \ \cdots \ , X_N^1, \cdots, X_N^{n_N}).$$

Suppose that $p_{iid(i)}$ is the variational form (as shown in Lemma 5) of p_{f_i}. Let the error caused by each atom as ϵ'_j $(1 \leq j \leq N)$. Then, total variation distance $\|p_{f_i} - p_{iid(i)}\|$ is bounded by $\sum_{j=1,\cdots,N} \epsilon'_j$.

Now, we prove that $\sum_{j=1,\cdots,N} \epsilon'_j \leq \sum_{i=1,\cdots,|F|} \epsilon_i$ ($|F|$ is the number of factors). For each factor f_i, ϵ_i is the sum of errors for all relational atoms included in the factor (i.e. $\epsilon_i =$

$\sum_j \text{s.t } {}_{X_j \in f_i} \epsilon'_j$). Given the fact that each relational atom is included in a factor at least once, $\sum_j \epsilon'_j \leq \sum_i \epsilon_i$. □

Appendix B. Analysis of the Lifted-MCMC Algorithm

In this section, we will present the details of the Lifted-MCMC algorithm. Then, we will show the correctness and computational complexity of the Lifted-MCMC algorithm.[9]

B.1 Correctness

We prove that our *Lifted-MCMC* converges to a correct stationary distribution, if a Gibbs sampling algorithm over a grounded RHM (e.g. [21]) converges to the stationary distribution.

Lemma 8. *If a Gibbs sampling algorithm over grounds variables in a RHM converges to a stationary distribution, the* **Lifted-MCMC** *algorithm converges to the stationary distribution.*

Proof. To prove the convergence, we prove that the *Lifted-MCMC* is *irreducible* which means that the Markov Chain can move between any pair of points. Then, we prove that *Lifted-MCMC* is ergodic.

We prove the *irreducibility* by contradiction. Assume that there is a sample $\mathscr{S}_L^0$=($I_{X_1} = d_1^0, \cdots, I_{X_N} = d_N^0$) which can not reach to another sample $\mathscr{S}_L^t$=($d_1^t, \cdots, d_N^t$). Suppose there is a map from each distribution d_i to a sample for corresponding RVs ($S_i = (X_i(a_1), \cdots, X_i(a_{|X_i|}))$). The map finds the most likely values for the RVs from the chosen distribution (e.g. d_i). In this way, $\mathscr{S}_L^0$ and $\mathscr{S}_L^t$ are mapped to $\mathscr{S}_G^0$=($S_1^0, \cdots, S_N^0$) and $\mathscr{S}_G^t$=($S_1^t, \cdots, S_N^t$), respectively.

The Gibbs sampler over ground RVs always finds a path from $\mathscr{S}_G^0$ and $\mathscr{S}_G^t$ because it is irreducible. (Otherwise, it can not converge to the stationary distribution.) WLOG, we assign the length of path as t so that we can refer a sample in a path as $\mathscr{S}_G^j$=($S_1^j, \cdots, S_N^j$) when $0 \leq j \leq t$. Now, for the jth sample in the path, we can define an inverse map which finds the most likely distribution d_i^j for each relational atom from the sample of ground RVs S_i^j.

In that way, we can prove that the *Lifted-MCMC* can move from $\mathscr{S}_L^j$ to $\mathscr{S}_L^{j+1}$ for all j. Suppose that, the ith latent variable d_i^j is changed to d_i^{j+1} by the inverse map. Then, the transition probability of the movement is determined by other latent variables. That is, the transition probability is positive whenever two samples $\mathscr{S}_G^j$ and $\mathscr{S}_G^{j+1}$ are reflected in finding potentials $\Phi()$. Because the factorization is exact, the transition probability between j and $j+1$ are positive for all j. It contradicts to the assumption. Thus, the *Lifted-MCMC* is *irreducible*.

Any finite state irreducible Markov Chain is *ergodic*. Based on the two properties (*irreducible* and *ergodic*), *Lifted-MCMC* converges to the stationary distribution. □

Thus, when Ground Gibbs sampling converges to a correct solution, *Lifted-MCMC* also converges to the correct solution.

9. Note that, the proof is about the convergence Lifted-MCMC algorithm is converged to the solution of a ground Gibbs sampling algorithms. The solution may incdlue the error caused by the variational form represented in Theorem 7.

B.2 Complexity

Now, we analyze the computational complexity and memory requirement of **Lifted-MCMC** algorithm because other two algorithms are one time batch procedures. We use Ω to refer a set of all ground RVs. For each relational atom X_i, $|X_i|$ refers the number of all ground RVs in the atom. That is, $U = \{X_i | X_i \in f, f \in F\}$ when F is a set of factors for a RHM.

Lemma 9. *The computational complexity of MCMC with ground RVs is* $O(n \cdot |\Omega|)$. *The space complexity is* $O(exp(\sum_{X_i \in F} |X_i|))$, *such that* $\arg\max_f \sum_{X_i \in F} |X_i|$. *($f$ is a factor that includes the largest number of ground RVs)*

Proof. The computational complexity is straightforward. The space complexity is determined by a factor that includes the largest number of ground RVs. That is, the size of CPT in the factor is exponentially proportional to **the number of all ground RVs** in the factor, g. □

Theorem 10. *The computational complexity of* Lifted-MCMC *is* $O(n \cdot |X|)$. *The complexity is* $O(exp(|\{X_i | X_i \in f\}|))$, *such that* $\arg\max_f |\{X_i | X_i \in f\}|$. *($f$ is a factor that includes the largest number of relational atoms)*

Proof. The computational complexity is also straightforward. The space complexity is determined by a factor that includes the largest number of relational RVs, because *Lifted-MCMC* does not generate samples for ground RVs. The size of CPT in the factored factor is exponentially proportional to **the number of relational atoms** in the factor, f. □

B.3 Accuracy

To solve inference problems, our lifted algorithm requires much less number of samples than previous Gibbs sampling algorithm over ground RVs because it runs on a smaller sampling space. Here, we provide a proof for the faster convergence of our algorithm, so that it provides a accurate sampling given a limited resource (e.g. limited number of samples).

Here, we calculate the total variation distance between the approximation ϕ_{approx} and the target distribution ϕ_{target} as follows.

$$\sum_{X(a_1),\cdots,X(a_n)} \left| \phi_{\text{target}}(X(a_1),\cdots,X(a_n)) - \phi_{\text{approx}}(X(a_1),\cdots,X(a_n)) \right|$$

Theorem 11. *After convergence to the stationary distribution, the distribution error of* Lifted-MCMC *is bounded by* $\frac{k}{N} + \sum_i \epsilon_i$ *when N is the number of samples,* ϕ_{target} *is factored by k mixtures, and* ϵ_i *is the error bound of each factor* f_i *in a RHM. After convergence, the error of any ground based Gibbs sampling is bounded by* $\frac{exp(n)}{N}$ *with n ground variables and N samples.*

Proof. Each sample at time t follows the stationary distribution because they are already converged. Thus, we focus on the error to estimate the ϕ_{target}.

When the factorization is exact, for each k distributions of latent variables, the error of density function is bounded by $\frac{1}{N}$ (i.e. $|\Phi_{\text{target}}(i) - \Phi_{\text{Lifted}}(i)| \leq \frac{1}{N}$). Thus, the error of

Lifted-MCMC over all possible values as follows.

$$\begin{aligned}
&\sum_{X(a_1),\cdots,X(a_n)} |\phi_{\text{target}}(X(a_1),\cdots,X(a_n)) - \phi_{\text{Lifted}}(X(a_1),\cdots,X(a_n))| \\
&= \sum_{X(a_1),\cdots,X(a_n)} \sum_{i=1,\cdots,k} \prod_{X(a_i)} \phi_{X_i}(X(a_i)) |\Phi_{\text{target}}(i) - \Phi_{\text{Lifted}}(i)| \\
&= \sum_{i=1,\cdots,k} |\Phi_{\text{target}}(i) - \Phi_{\text{Lifted}}(i)| \sum_{X(a_1),\cdots,X(a_n)} \prod_{X(a_i)} \phi_{X_i}(X(a_i)) \\
&= \sum_{i=1,\cdots,k} |\Phi_{\text{target}}(i) - \Phi_{\text{Lifted}}(i)| \leq \frac{k}{N}
\end{aligned}$$

When the factorization is not exact, the error is an addition of $\frac{k}{N}$ and $\sum_i \epsilon_i$ in Theorem 7.

In the ground case, the error of density function is also bounded by $\frac{1}{N}$ for each value of RVs. $|\phi_{\text{target}}(x_1,\cdots,x_n) - \Phi_{\text{Ground}}(x_1,\cdots,x_n)| \leq \frac{1}{N}$. The error of *Ground-MCMC* over all possible values $exp(n)$ is as follows,

$$\begin{aligned}
&\sum_{X(a_1),\cdots,X(a_n)} |\phi_{\text{target}}(X(a_1),\cdots,X(a_n)) - \phi_{\text{Ground}}(X(a_1),\cdots,X(a_n))| \\
&\leq \sum_{X(a_1),\cdots,X(a_n)} \frac{1}{N} = \frac{exp(n)}{N} \quad \square
\end{aligned}$$

Theorem 11 shows how the cardinality of sampling space affects the error of the posterior distribution. For continuous cases, a similar proof can be applied.

References

[1] B. Ahmadi, K. Kersting, and S. Sanner. Multi-evidence lifted message passing, with application to pagerank and the kalman filter. In *IJCAI*, 2011.

[2] Jaesik Choi, Abner Guzman-Revera, and Eyal Amir. Lifted relational kalman filtering. In *IJCAI*, 2011.

[3] Jaesik Choi, David Hill, and Eyal Amir. Lifted inference for relational continuous models. In *UAI*, 2010.

[4] Wei Chu, Vikas Sindhwani, Zoubin Ghahramani, and S. Sathiya Keerthi. Relational learning with gaussian processes. In *NIPS*. AAA, 2006.

[5] Sanjoy Dasgupta. Learning mixtures of gaussians. In *FOCS*, 1999.

[6] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Mathematice e Naturale*, 0, 1931.

[7] Rodrigo de Salvo Braz, Eyal Amir, and Dan Roth. Lifted first-order probabilistic inference. In *IJCAI*, 2005.

[8] P. Diaconis. Finite forms of de finetti's theorem on exchangeability. *Synthese*, 0, 1977.

[9] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Annals of Probability*, 0, 1980.

[10] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, 1999.

[11] Abhay Jha, Vibhav Gogate, Alexandra Meliou, and Dan Suciu. Lifted inference seen from the other side : The tractable features. In *NIPS*, 2010.

[12] J. F. C. Kingman. Uses of Exchangeability. *The Annals of Probability*, 6(2):183–197, April 1978.

[13] Brian Milch and Stuart J. Russell. First-order probabilistic languages: Into the unknown. In *ILP*, 2006.

[14] Raymond Ng and V. S. Subrahmanian. Probabilistic logic programming. *Inf. Comput.*, 101(2), 1992.

[15] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart J. Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *NIPS*, 2002.

[16] Avi Pfeffer, Daphne Koller, Brian Milch, and Ken T. Takusagawa. Spook: A system for probabilistic object-oriented knowledge representation. In *UAI*, 1999.

[17] David Poole. First-order probabilistic inference. In *IJCAI*, 2003.

[18] Carl E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *NIPS*, 2002.

[19] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2), 2006.

[20] Parag Singla and Pedro Domingos. Lifted first-order belief propagation. In *AAAI*, 2008.

[21] Jue Wang and Pedro Domingos. Hybrid markov logic networks. In *AAAI*, 2008.

[22] Lei Xu and Michael I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, 8:129–151, 1995.

[23] Zhao Xu, Kristian Kersting, and Volker Tresp. Multi-relational learning with gaussian processes. In *IJCAI*, 2009.