Lifted Hybrid Variational Inference

Yuqiao Chen[*1], **Yibo Yang**[*2], **Sriraam Natarajan**[1] and **Nicholas Ruozzi**[1]
[1] The University of Texas at Dallas
[2] University of California Irvine
{yuqaio.chen, sriraam.natarajan, nicholas.ruozzi}@utdallas.edu, yibo.yang@uci.edu

Abstract

Lifted inference algorithms exploit model symmetry to reduce computational cost in probabilistic inference. However, most existing lifted inference algorithms operate only over discrete domains or continuous domains with restricted potential functions. We investigate two approximate lifted variational approaches that apply to domains with general hybrid potentials, and are expressive enough to capture multi-modality. We demonstrate that the proposed variational methods are highly scalable and can exploit approximate model symmetries even in the presence of a large amount of continuous evidence, outperforming existing message-passing-based approaches in a variety of settings. Additionally, we present a sufficient condition for the Bethe variational approximation to yield a non-trivial estimate over the marginal polytope.

1 Introduction

Probabilistic inference in Markov random fields (MRFs) and their generalizations are intractable in all but the simplest cases [Koller and Friedman, 2009]. Lifted inference methods tackle this challenge by exploiting symmetries in a given model. Such methods construct groups of indistinguishable random variables that can be used to collapse the model into a simpler one on which inference is more tractable.

High-level approaches to lifted inference include message-passing algorithms such as lifted belief propagation (BP) [Singla and Domingos, 2008; Kersting *et al.*, 2009] and lifted variational methods [Bui *et al.*, 2014; Gallo and Ihler, 2018]. The common theme across these methods is the construction of a lifted graph on which the corresponding inference algorithms are run. The message-passing algorithms are applied directly on the lifted graph, while lifted variational methods encode symmetries in the model as equality constraints in the variational objective. These two approaches are directly related via the same variational objective, the Bethe free energy [Yedidia *et al.*, 2001; Yedidia *et al.*, 2005]. While successful, these methods were designed for discrete relational MRFs.

Existing work on lifting with continuous domains has focused primarily on Gaussian graphical models [Choi *et al.*, 2010; Choi *et al.*, 2011; Ahmadi *et al.*, 2011]. Hybrid Lifted BP [Chen *et al.*, 2019] extends lifted belief propagation to continuous domains by using particle message passing and coarse-to-fine approximate lifting to handle continuous evidence. With this method, the quality of inference largely depends on sampling. While it performs well on models with simple marginals, it can fail to accurately model multi-modal distributions and may have numerical or convergence issues. Another lifted inference method for generic hybrid domains [Choi and Amir, 2012] uses expectation maximization to learn variational approximations of the MRF potentials (this requires sampling from the potentials, which implicitly assumes the potentials are normalizable) and then performs lifted variable elimination or MCMC on the resulting variational model.

Our aim is to provide a general framework for lifted variational inference that can be applied to both continuous and discrete domains with generic potential functions. Our approach is based on mixtures of mean-field models and a choice of entropy approximation. We consider two entropy approximations, one based on the Bethe free energy, whose local optima are closely related to fixed points of BP [Yedidia *et al.*, 2001], and a lower bound on the differential entropy based on Jensen's inequality [Gershman *et al.*, 2012].

We make the following key contributions: (1) We develop the **first generic lifted hybrid variational approach** for probabilistic inference: our approach does not make any distributional or model assumptions, and it can be applied to arbitrary factor graphs. (2) We consider two different types of approximations based on mixtures of mean-field models. To our knowledge, **a systematic comparison of these two different approximations for continuous models does not exist** in the literature. (3) We provide **theoretical justification for the Bethe free energy in the continuous case** by providing a sufficient condition for it to be bounded from below over the marginal polytope. (4) We demonstrate the superiority of our approach empirically against particle-based message-passing algorithms and variational mean field.

2 Preliminaries

A Markov Random Field (MRF) is specified by a hypergraph $G = (\mathcal{V}, \mathcal{C})$ with node set $\mathcal{V}$ and hyper-edge/clique

[*]denotes equal contributions.

set $\mathcal{C}$. Each node $i \in \mathcal{V}$ corresponds to a random variable x_i with domain $\mathcal{X}_i$, and each clique c in $\mathcal{C}$ is associated with a non-negative potential function $\psi_c : \prod_{i \in c} \mathcal{X}_i \to \mathbb{R}_{\geq 0}$ defined over $x_c = \{x_i; \forall i \in c\}$. A given MRF defines a joint probability distribution $p(x) = \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \psi_c(x_c)$ over $x = \{x_i; \forall i \in \mathcal{V}\}$, where $\mathcal{Z}$ is a normalizing constant.

We consider *hybrid* MRFs, i.e., MRFs with both discrete and continuous random variables, so that $\mathcal{X}_i$ may either be finite or uncountable. If all of the variables have continuous domains and the product of potential functions is integrable, then the normalization constant exists, e.g., $\mathcal{Z} \triangleq \int_{x \in \prod_{i \in \mathcal{V}} \mathcal{X}_i} \prod_{c \in \mathcal{C}} \psi_c(x_c) < \infty$. The hypergraph G is often visualized as a factor graph that represents cliques as factor nodes and variables as variable nodes, with an edge joining the factor node c to the variable node i if $i \in c$.

We consider two *probabilistic inference* tasks for a given MRF: (1) marginal inference, i.e., computing the marginal probability distribution $p(x_\mathcal{A})$ of a set of variables $\mathcal{A} \subseteq \mathcal{V}$, a special case of which is computing the partition function $\mathcal{Z}$ when $\mathcal{A} = \mathcal{V}$; and (2) maximum a posteriori (MAP) inference, i.e., computing $\arg\max_{x_\mathcal{A}} p(x_\mathcal{A})$ of the distribution $p(x_\mathcal{A})$. In many applications, we will be given observed values $\mathbf{x}_\mathcal{B}$ for a set of variables $\mathcal{B} \subseteq \mathcal{V}$, and the corresponding conditional marginal / MAP inference tasks involve computing $p(x_\mathcal{A}|\mathbf{x}_\mathcal{B})$ instead of $p(x_\mathcal{A})$.

Variational Inference (VI) solves the inference problem approximately by minimizing some divergence measure D, often chosen to be the Kullback-Leibler divergence, between the true model and a family of more tractable approximate distributions $\mathcal{Q}$, to obtain a surrogate distribution $q^* \in \arg\min_{q \in \mathcal{Q}} D(q, p)$. The set $\mathcal{Q}$ is typically chosen to trade off between the computational ease of inference in a surrogate model q and its ability to model complex distributions. When D is the KL divergence, the optimization problem is equivalent to minimizing the variational free energy:

$$\mathcal{F}(q) = -\sum_{c \in \mathcal{C}} \mathbb{E}_q[\psi_c] - \mathbb{H}[q] \tag{1}$$

where $\mathbb{H}[q]$ denotes the entropy of the distribution q. Assuming one can find a $q^* \in \arg\min_q \mathcal{F}(q)$, the *simpler* model q^* can be used as a surrogate for inference. A popular choice for $\mathcal{Q}$ is the set of completely factorized distributions, a.k.a. mean-field approximation, which greatly simplifies the optimization problem.

Lifted Inference exploits symmetries that exist in the MRF in order to reduce the complexity of inference. This is typically done by grouping symmetric variables or cliques together into a single super variable/clique and then tying together the corresponding marginals of all variables in the same super variable/clique [Bui *et al.*, 2014]. Detecting symmetries can be done in a top-down [Singla and Domingos, 2008] or bottom-up [Kersting *et al.*, 2009] fashion. We use the color passing (CP) algorithm [Kersting *et al.*, 2009], a bottom-up approach that can be applied to arbitrary MRFs. In CP, all variable and factor nodes are initially clustered based on domain/evidence and the potential functions. Variables with the same domain or the same evidence value v will be assigned the same color. Each clique node stacks the color of its neighboring nodes in order and appends its own color. Each variable node collects the colors of its neighboring cliques and is assigned a new color. The process is repeated until convergence. The color information is the neighborhood structure information and grouping nodes with the same color can be used to compress the graph. We use the notation $\#(\mathfrak{c})$ and $\#(\mathfrak{i})$ to denote the number of factors in a super factor $\mathfrak{c}$ and the number of variables in super variable $\mathfrak{i}$, respectively.

3 Proposed Approaches

Our aim is to develop distribution-independent, model-agnostic lifted variational inference algorithms that operate on arbitrary hybrid MRFs. To overcome the limitations of unimodal variational distributions, e.g., mean-field, we choose our approximate family $\mathcal{Q}$ to be a family of mixture distributions, and following Jaakkola and Jordan [1998] and Gershman et al. [2012], we require each mixture component to be fully factorized. Specifically,

$$q(x) = \sum_{k=1}^{K} w_k q^k(x) = \sum_{k=1}^{K} w_k \prod_i q_i^k(x_i|\eta_i^k), \tag{2}$$

where K is the number of mixture components, $w_k \geq 0$ is the weight of the k^{th} mixture (a shared parameter across all marginal distributions), and $\sum_{k=1}^{K} w_k = 1$. Each $q_i^k(x_i) \triangleq q_i^k(x_i; \eta_i^k)$ is some valid univariate distribution with parameters η_i^k, e.g., a Gaussian or Beta distribution in the continuous case, or a Categorical distribution in the discrete case.

3.1 Entropy Approximations

Ideally, we would find the appropriate model parameters η and w by directly minimizing the VI objective Eq (1). Unfortunately, computing the entropy $\mathbb{H}[q]$ is intractable for arbitrary variational distributions of the form (2). A notable exception is when K=1, which is equivalent to the naïve mean-field approximation. In the general case, we consider two tractable entropy approximations: one based on Bethe free energy approximation and one based on Jensen's inequality.

The Bethe Entropy. $\mathbb{H}_B$ approximates $\mathbb{H}$ as if the graph G associated with p were tree-structured:

$$\mathbb{H}_\mathrm{B}[q] \triangleq \sum_{c \in \mathcal{C}} \mathbb{H}[q_c] + \sum_{i \in \mathcal{V}} (1 - |nb(i)|)\mathbb{H}[q_i],$$

where $nb(i) = \{c \in \mathcal{C} | i \in c\}$ is the set of cliques that contain node i in their scope. The Bethe free energy (BFE) is then defined as

$$\mathcal{F}_\mathrm{B}(q) \triangleq -\sum_c \mathbb{E}_q[\psi_c] - \mathbb{H}_B[q].$$

The BFE approximation is exact whenever the hypergraph G is acyclic, i.e., tree-structured. While variational methods seek to optimize the variational objective directly, message-passing algorithms such as belief propagation (BP) can also be used to find local optima of the BFE [Yedidia *et al.*, 2001]. As message-passing algorithms can suffer from convergence issues, gradient-based methods that optimize the variational objective directly are sometimes preferred [Welling and Teh, 2001; Guo *et al.*, 2019a].

Jensen's Inequality. Non-parametric variational inference (NPVI) approximates the entropy using Jensen's inequality [Gershman *et al.*, 2012].

$$\mathbb{H}_J[q] = -\sum_k w_k \log(\sum_j w_j \mathbb{E}_{q^k}[q^j]) \tag{3}$$

There are two reasons to prefer the Bethe entropy approximation over the NPVI lower bound approximation (3). First, the BFE is exact on trees; for tree-structured models, it is likely to outperform NPVI. Second, the NPVI lower bound (3) does not factorize over the graph, potentially making it less useful in distributed settings. Conversely, one advantage of the NPVI approximation over the Bethe entropy is that it gives rise to a provable lower bound on the partition function $\mathcal{Z}$ assuming exact computation. The Bethe entropy only provably translates into a lower bound on tree structured models or for special model classes [Ruozzi, 2012; Ruozzi, 2013; Ruozzi, 2017].

Another known drawback of the BFE is that, in the case of continuous random variables, it need not be bounded from below over the local marginal polytope. The local marginal polytope is a further relaxation of the variational problem in which the optimization over distributions $q \in \mathcal{Q}$ is replaced by a simpler optimization problem over only marginal distributions that agree on their univariate marginals. Unboundedness can occur even in Gaussian MRFs [Cseke and Heskes, 2011]. This makes BFE potentially undesirable for continuous MRFs in practice. However, for the optimization problem considered here (over a subset of marginals (2) that are globally consistent, referred to as the *marginal polytope*), it is known that the BFE is bounded from below for Gaussian MRFs [Guo *et al.*, 2019b]. Here, we generalize this result to a larger class of distributions induced by MRFs, further justifying the use of BFE in our variational framework.

Theorem 1. *If there exists a collection of densities $g_i \in \mathcal{P}(\mathcal{X}_i)$ for each $i \in \mathcal{V}$ such that $\sup_{x\in\mathcal{X}} \frac{p(x)}{\prod_i g_i(x_i)} < \infty$, then $\inf_{q\in\mathcal{P}(\mathcal{X})} -\mathbb{E}_q[\log p] - \mathbb{H}_{\mathrm{B}}[q] > -\infty$, where $\mathcal{P}(\mathcal{X})$ is the set of all probability densities over $\mathcal{X} \triangleq \prod_{i\in\mathcal{V}} \mathcal{X}_i$, i.e., the BFE is bounded from below.*

Many naturally occurring distributions satisfy the condition of the theorem: mean-field models, multivariate Gaussians and their mixtures, bounded densities with compact support, etc. The proof of the theorem is given in Appendix A.

3.2 Lifted Variational Inference

Once symmetries are detected using CP or an alternative method, they can be encoded into the variational objective by introducing constraints on marginals, e.g., adding a constraint that all variables in the same super node have equivalent marginals. This is the approach taken by [2014] for lifted variational inference in discrete MRFs. In our mixture distribution setting, this leads to the following,

$$\forall i, j \in \mathfrak{i}, \sum_{k=1}^{K} w_k q_i^k(x_i) = \sum_{k=1}^{K} w_k q_j^k(x_j) \tag{4}$$

If preferred, these constraints could be incorporated into the objective as a soft penalty to encourage the solution to contain the appropriate symmetries as discovered by color-passing. However, adding constraints of this form to the objective does not reduce the cost of performing inference in the lifted model. In order to make inference efficient, we observe that the following constraints are sufficient for Eq (4) to hold.

$$\forall i, j \in \mathfrak{i}, \forall_k, q_i^k(x_i) = q_j^k(x_j) \tag{5}$$

Under the constraints in Eq (5), we can simplify the variational objective by accounting for the shared parameters. Consider a compressed graph $\mathfrak{G}$ with a set of super variables $\mathfrak{V}$ and a set of super factors $\mathfrak{C}$, each $\mathfrak{i} \in \mathfrak{V}$ and each $\mathfrak{c} \in \mathfrak{C}$ corresponds to $\#(\mathfrak{i})$ variables and $\#(\mathfrak{c})$ factors in the original graph. Variables in the super variable $\mathfrak{i}$ share the same parameterized marginals. Using this observation, we simplify the computation of unlifted BFE by exploiting these symmetries:

$$\sum_{\mathfrak{i}\in\mathfrak{V}} \#(\mathfrak{i}) \cdot \mathbb{E}_{q_{\mathfrak{i}}(x_{\mathfrak{i}})}\Big[(1 - |nb(\mathfrak{i})|)\log q_{\mathfrak{i}}(X_{\mathfrak{i}})\Big] + \sum_{\mathfrak{c}\in\mathfrak{C}} \#(\mathfrak{c}) \cdot \mathbb{E}_{q_{\mathfrak{c}}(x_{\mathfrak{c}})}\Big[\log q_{\mathfrak{c}}(X_{\mathfrak{c}}) - \log \psi_{\mathfrak{c}}(X_{\mathfrak{c}})\Big].$$

The NPVI approximation Eq (3) can be lifted similarly using parameter sharing conditions (5). Concretely, the innermost expectation in Eq (3) can be re-written in terms of the variational parameters associated with underlying super variables:

$$\begin{aligned}\mathbb{E}_{q^k}[q^j] &= \prod_{i\in\mathcal{V}} \int q_i^k(x_i) q_i^j(x_i) dx_i \\ &= \prod_{\mathfrak{i}\in\mathfrak{V}} [\int q_{\mathfrak{i}}^k(x_{\mathfrak{i}}) q_{\mathfrak{i}}^j(x_{\mathfrak{i}}) dx_{\mathfrak{i}}]^{\#(\mathfrak{i})}\end{aligned}$$

Although the optimal value of the variational objective under the constraints (4) or (5) is always greater than or equal to that of the unconstrained problem, we expect gradient descent on the constrained optimization problem to converge faster and to a better solution as the optimal solution should contain these symmetries. The intuition for this is that the solutions to the unconstrained optimization problem, i.e., approximate inference in the unlifted model, can include both solutions that do and do not respect the model symmetries.

Given a ground MRF (possibly with evidence) and a choice of entropy approximation (either using Bethe entropy or NPVI), we first find the variational distrubtion q^* by gradient descent on the appropriate VI objective $\mathcal{F}(q)$ (Eq (1)) w.r.t. the parameters (w, η) of mean-field variational mixture q, where the kth component q_i^k of variable marginal is taken to be a Gaussian distribution for all continuous variable $i \in \mathcal{V}$ and a categorical distribution otherwise. The lifted VI algorithms additionally exploit symmetries by using (5) to simplify the objective and only optimize over the variational parameters $\eta_{\mathfrak{i}}$ associated with the super variables $\mathfrak{i} \in \mathfrak{V}$; after the optimization procedure, all the original variables contained in each super variable are assigned the same variational marginal parameters as in (5). The expectations in the variational objectives can be approximated in several ways – sampling, Stein variational methods [Liu and Wang, 2016], etc. We approximate the expectations using Gaussian quadrature [Golub and Welsch, 1969]. Once q^* is obtained, given a set of query variables U, marginal inference is approximated by

$p(x_U) \approx q^*(x_U) = \sum_{k=1}^{K} w_k \prod_{i \in U} q_i^{*k}(x_i)$, and (marginal) MAP is approximated by $\arg\max_{x_U} q^*(x_U)$ via coordinate ascent or gradient ascent.

3.3 Coarse-to-Fine Lifting

A common issue in lifting is that introducing evidence breaks model symmetries as variables with different evidence values should be considered as different even if they have similar neighborhood structure. This issue is magnified when variables are continuous: it is unlikely that two otherwise symmetric variables will receive the exact same evidence values. As a result, even with a small amount of evidence, many of the model symmetries may be destroyed, making lifting less useful. To counteract this effect, we propose a coarse-to-fine (C2F) approximate lifting method in the variational setting that is based on the assumption that the stationary points of a coarsely compressed graph and a finely compressed graph should be similar. A number of C2F lifting schemes, which start with coarse approximate symmetries and gradually refine them, have been proposed for discrete MRFs [Habeeb *et al.*, 2017; Gallo and Ihler, 2018]. Our approach specifically aims to approximate symmetries to handle the above issue with continuous evidence.

Our C2F approximate lifting uses k-means clustering to group the continuous evidence values into s clusters, $E_1, \ldots, E_s$. For each cluster $E_i \in \{E_1, \ldots, E_s\}$, we denote the corresponding set of observation nodes as O_i. Each observed variable $o \in O_i$ is treated as having the same evidence distribution $b_{E_i}(x_o) = \mathcal{N}(\mu_{E_i}, \sigma^2_{E_i})$, where μ_{E_i} and $\sigma^2_{E_i}$ are the mean and variance of cluster E_i. With this formalism, the evidence clustering is coarse when s is small, but we can exploit more approximate symmetries, resulting in a more compressed lifted graph. As s increases, the evidence variables are more finely divided.

To apply this lifting process in VI, we interleave the operation of refining the compressed graph and gradient descent. The clustering is initialized with $s = 1$ and CP is run until convergence to obtain a coarse compressed graph. Then, we perform gradient descent on the coarse compressed graph with the lifted Bethe variational method. After a fixed number of iterations, we refine the coarse compressed graph by splitting evidence clusters. We use the k-means algorithm to determine the new evidence clusters, and obtain a refined compressed graph using CP. We keep iterating this process until no evidence group can be further split, e.g., when only one value remains or the variance of each cluster is below a specified threshold, and the optimization converges to a stationary point. A precise description of this process can be found in Algorithm 1. CP is not run from the start after each split: we simply assign a new evidence group and a new color and resume CP from its previous stopping point.

Algorithm 1 Coarse-to-Fine Lifted VI

Input: A factor graph G, evidence E and splitting threshold ϵ
Return: The model parameters η and w
$\mathcal{E}, \eta, w \leftarrow$ initial clustering of continuous evidence and model parameters respectively
Group variables with same domain/evidence distribution and factors with same potential together
$\mathfrak{G} \leftarrow$ run CP until convergence
repeat
    $\eta, w \leftarrow$ run grad. descent on variational obj.
    for each $E_i \in \mathcal{E}$ **do**
        if $\sigma^2_{Ei} > \epsilon$ **then**
            $\mathcal{E}_i \leftarrow$ Divide E_i in two using k-means
            $\mathcal{E} \leftarrow (\mathcal{E} \setminus E_i) \cup \mathcal{E}_i$
        end if
    end for
    Assign new colors to evidence according to $\mathcal{E}$
    $\mathfrak{G} \leftarrow$ run CP until convergence
until convergence

4 Experiments

We investigate the performance of the proposed lifted variational inference approach on a variety of both real and synthetic models. We aim to answer the following questions explicitly: **Q1:** Do the proposed variational approaches yield accurate MAP and marginal inference results? **Q2:** Does lifting result in significant speed-ups versus an unlifted variational method? **Q3:** Does C2F lifting yield accurate results more quickly for queries with continuous evidence?

We compare the performance of our variational approach using different entropy approximations (denoted "BVI" for Bethe approximation and "NPVI" for Jensen's lower bound approximation) with message-passing algorithms including Expectation Particle BP (EPBP) [Lienart *et al.*, 2015], Hybrid Lifted BP (HLBP) [Chen *et al.*, 2019], and Gaussian BP (GaBP). To illustrate the generality of our results, we consider three different model settings – Hybrid Markov Logic Networks (HMLNs) [Wang and Domingos, 2008], Relational Gaussian Models (RGMs) [Choi *et al.*, 2010], and Relational Kalman Filters (RKFs) [Choi *et al.*, 2011].

We report the ℓ_1 error of MAP predictions and KL divergence $D_{KL}(p_i||q_i)$ averaged across all univariate marginals i in the (ground) conditional MRF. As the models in the RGM and RKF experiments are Gaussian MRFs, their marginal means and variances can be computed exactly by matrix operations. For the HMLN experiments, the exact ground truth can be obtained by direct methods when the number of random variables in the conditional MRF is small. All timing results, unless otherwise noted, were performed on a single core of a 2.2 GHz Intel Core i7-8750H CPU with 16GB memory. Source code is available on https://github.com/leodd/Lifted-Hybrid-Variational-Inference.

4.1 Hybrid MLNs

Unlike standard MLN, where only first order formulas are allowed, HMLN extends it by adding continuous formulas and hence its grounded model corresponds to a MRF with hybrid domain. In this section, we first consider a toy HMLN with known ground truth marginals in order to assess the accuracy of the different variational approaches in the hybrid setting. Then, we showcase the efficiency of our methods, particularly via lifting, on larger-scale HMLNs of practical interest.

We construct a **Toy Hybrid MLN** for a *position* domain:

$$0.1 : in(A, Box) \wedge in(B, Box) \rightarrow attractedTo(A, B)$$
$$0.2 : \neg attractedTo(A, B) \cdot [pos(A) = p1] +$$
$$attractedTo(A, B) \cdot [pos(B) = p2],$$

where A and B are different classes of objects in a physics simulation, Box is the class of box instances, and $p1, p2$ are real values corresponding to object positions. Predicates in and $attractedTo$ have discrete domain $\{0, 1\}$, while pos is real-valued. The equality operation $x = y$, is the shorthand for $-(x-y)^2$ which corresponds to linear Gaussian potential $\exp(-(x-y)^2)$ in the grounded MRF.

The marginals of $pos(A)$ and $pos(B)$ will generally be multimodal, specifically mixtures of Gaussians; with multiple object instances, unimodal variational approximations like mean-field will likely be inaccurate. Note that the number of mixture components in the marginal distributions is exponential in the number of joint discrete configurations in the ground MRF, so we consider a small model in which exact inference is still tractable, generating 2, 3, and 2 instances of the A, B, and Box classes respectively. The resulting ground MRF contains 16 discrete random variables yielding marginals with 2^{16} mixture components. The small model size is only for the purpose of evaluation against brute-force exact inference; our methods can scale to larger models.

We performed marginal inference on the continuous nodes and report results against ground truth in Table 1, using BVI, NPVI, and their lifted versions (abbreviated as L-BVI and L-NPVI). All methods tend to give improved performance as the number of mixture components (K) increases indicating that the number of mixture components is indeed important for accuracy in multimodal settings. However, increasing K generally makes the optimization problem more difficult, requiring more iterations for convergence. We note that even though L-BVI reported a lower ℓ_1 error with $K = 1$, the KL-divergence of this was larger than at $K = 3$ or 5, indicating that it converged to a good local optimum for the MAP task but not as good for the marginal inference task. This distinction can be seen more broadly across the two entropy approximations for this problem, as BVI/L-BVI generally gave better fits to the marginals than NPVI/L-NPVI, whereas NPVI/L-NPVI performed better at estimating the marginal modes.

It is also worth noting that *lifting seems to act as a regularizer* here: when the number of mixture components is small, both the lifted versions outperformed their unlifted counterparts, e.g., at K=1. This suggests that lifting may both reduce computational cost (30% to 40% speedup on this model) and encourage the optimization procedure to end up in better local optima, which positively answers **Q1** and **Q2**.

Next, we consider two larger scale HMLNs of practical interest. The **Paper Popularity** HMLN domain is determined by the following formulas.

$$0.3 : PaperPopularity(p) = 1.0$$
$$0.5 : SameSession(t1, t2) \cdot$$
$$[TopicPopularity(t1) = TopicPopularity(t2)]$$
$$0.5 : In(p, t) \cdot [PaperPopularity(p) = TopicPopularity(t)],$$

Algorithm	Average KL-Divergence		
	$K = 1$	$K = 3$	$K = 5$
BVI	0.513 ± 0.947	0.009 ± 0.006	0.009 ± 0.007
L-BVI	$0.039 \pm 1\text{e}-5$	$\mathbf{0.004 \pm 0.001}$	0.004 ± 0.002
NPVI	4.586 ± 2.977	0.026 ± 0.007	0.022 ± 0.003
L-NPVI	1.978 ± 3.878	0.038 ± 0.002	0.039 ± 0.001
Algorithm	Average ℓ_1-error		
	$K = 1$	$K = 3$	$K = 5$
BVI	0.227 ± 0.440	0.059 ± 0.026	0.049 ± 0.035
L-BVI	$0.007 \pm 1\text{e}-4$	0.049 ± 0.027	0.020 ± 0.012
NPVI	1.641 ± 1.106	$\mathbf{0.007 \pm 4\text{e}-4}$	0.012 ± 0.009
L-NPVI	0.671 ± 1.327	0.012 ± 0.006	0.012 ± 0.006

Table 1: Results of variational methods on toy HMLN.

where $PaperPopularity(p)$ and $TopicPopularity(t)$ are continuous variables in $[0, 10]$, indicating the popularity of paper and topic. $SameSession(t1, t2)$ and $In(p, t)$ are Boolean variables, indicating if two topics are in the same session, and if a paper p belongs to a topic t, respectively. We instantiate 300 paper instances and 10 topic instances, which results in $3{,}400$ variables and $3{,}390$ factors in the grounded MRF. We generated random evidence for the model, where 70% of the papers and 70% of the topics were assigned a popularity from a uniform distribution $U(0, 10)$. For 70% of the papers p and all topics t, we assign $In(p, t')$ and $SameSession(t, t')$ for all possible $t' \in Topic$ using a Bernoulli distribution ($p = 0.5$).

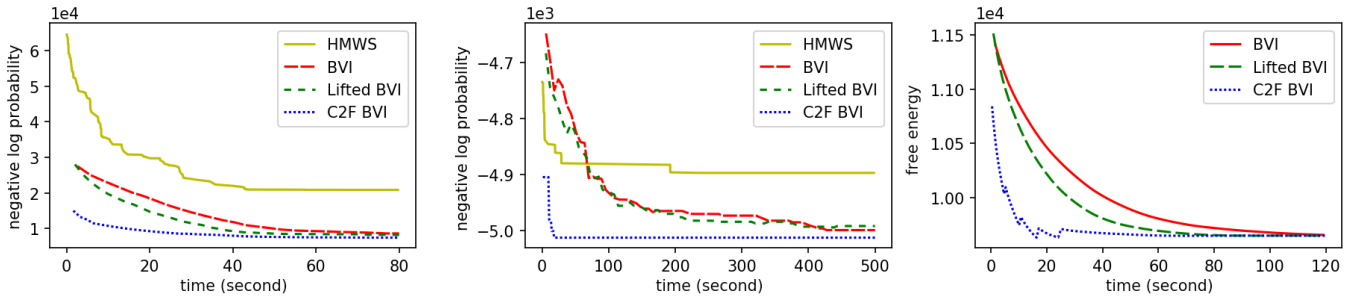
The **Robot Mapping** HMLN domain contains 3 discrete relational variables, 2 continuous relational variables, and 10 formulas, as described in the Alchemy tutorial [Wang and Domingos, 2008]. The instances and evidences are from real world robot scanning data, which result in a grounded MRF with $1{,}591$ random variables and $3{,}182$ factors.

Results: We performed MAP inference with the VI methods and evaluated them against Hybrid MaxWalkSAT (HMWS) [Wang and Domingos, 2008]. Each method is evaluated by computing the energy $E(\hat{x}) = -\log(\prod_{c \in \mathcal{C}} \psi_c(\hat{x_c}))$, essentially the negative log probability, of its estimated MAP configuration. For HMWS, we set the greedy probability to 0.7, the standard deviation of the Gaussian noise to 0.3, and disabled re-running for fair comparison.

As can be seen in Figure 1, in both domains, the MAP assignment produced by the VI methods is significantly better than the one by HMWS, providing support for **Q1**. In addition, given the amount of continuous evidence, there is not a significant performance difference between BVI and Lifted BVI. However, C2F BVI takes significantly less time to converge to a good solution than both BVI and Lifted BVI, providing strong evidence for **Q3**, namely that C2F results in better accuracy more quickly.

4.2 Relational Gaussian Models (RGMs)

We performed approximate inference on a RGM with the recession domain from [Cseke and Heskes, 2011]. The RGM has three relational atoms $Market(S)$, $Loss(S, B)$, $Revenue(B)$ and one random variable $Recession$, where S and B denote two sets of instances, the categories of market and banks respectively. For testing, we generated 100 Market and 5 Bank instances, and used the ground graph as input.



(a) Paper Popularity (b) Robot Mapping (c) RGM

Figure 1: (a) - (b) comparison of the negative log probability of the approximate MAP assignment versus running time. (c) comparison of rate of convergence between BVI, L-BVI, and C2F-BVI with 20% evidence.

Algorithm	Avg. ℓ_1 Error	Avg. KL-Divergence
EPBP	$5.17\text{e}{-2} \pm 4.25\text{e}{-2}$	0.473 ± 0.246
HLBP	$6.26\text{e}{-2} \pm 4.29\text{e}{-2}$	0.485 ± 0.259
BVI	$6.48\text{e}{-5} \pm 2.57\text{e}{-4}$	$4.95\text{e}{-3} \pm 5.13\text{e}{-2}$
L-BVI	$5.77\text{e}{-5} \pm 2.47\text{e}{-4}$	$4.95\text{e}{-3} \pm 5.13\text{e}{-2}$
C2F-BVI	$7.92\text{e}{-4} \pm 2.31\text{e}{-3}$	$4.95\text{e}{-3} \pm 5.13\text{e}{-2}$
NPVI	$3.22\text{e}{-5} \pm 2.89\text{e}{-5}$	$5.14\text{e}{-4} \pm 4.14\text{e}{-5}$
L-NPVI	$\mathbf{3.29e{-5} \pm 5.58e{-5}}$	$\mathbf{5.14e{-4} \pm 4.14e{-5}}$

Table 2: Evaluation of various methods on RGM.

To assess the impact of lifting and C2F, we randomly chose 20% of the variables, assigned them a value uniformly randomly from $[-30, 30]$, and then performed conditional MAP and marginal inference.

Figure 1c plots the variational free energy, i.e., the VI objective Eq (1), versus CPU time for BVI, L-BVI, and C2F-BVI. All three methods used the Adam optimizer with learning rate $0.2, \beta_1 = 0.9, \beta_2 = 0.999$. The plot shows that C2F-BVI converges faster than L-BVI, which is in turn faster than BVI. Note that the sawtooth shape of C2F-BVI is a result of evidence splitting. This shows that lifting and C2F reduces cost of inference, answering **Q2** and **Q3** affirmatively.

To assess the accuracy of inference methods, we randomly chose 5% to 20% of the random variables and generated evidence values as in the previous task. We randomly generated five evidence settings and evaluated all VI methods with the same setup as above as well as EPBP with 20 sampling points. All algorithms were run to convergence and compared against the ground truth. As Table 2 shows, on this simple unimodal model, all VI methods have very low error and KL-divergence, while particle-based methods have higher error / KL-divergence as a result of the sampling procedure, providing evidence for **Q1**. In general, NPVI tends to estimate the mode of the distribution better than BVI but in multimodal settings tends to result in a higher KL-divergence than BVI.

4.3 Relational Kalman Filtering

To further investigate **Q1**, we performed an experiment with Relational Kalman Filters (RKFs). A standard Kalman filter (KF) models the transition of a dynamic system with $x_{t+1} = Ax_t + w$ with noise $o_t = Cx_t + v$, where A denotes the transition matrix and C represents the observation matrix. A key assumption in KF is that the transition and noise follow

	Algorithm	ℓ_1 Diff.(GaBP)	time (s)
Tree	EPBP	0.18 ± 0.15	233.6
	HLBP	0.25 ± 0.24	**8.0**
	BVI	$2.86\text{e}{-5} \pm 3.71\text{e}{-5}$	577.2
	L-BVI	$\mathbf{1.88e{-5} \pm 1.85e{-5}}$	23.0
	C2F-BVI	$1.64\text{e}{-5} \pm 1.80\text{e}{-5}$	**22.8**
Cycle	EPBP	0.34 ± 0.41	149.2
	HLBP	0.36 ± 0.55	**75.9**
	BVI	$\mathbf{3.26e{-5} \pm 5.08e{-5}}$	222.1
	L-BVI	$4.65\text{e}{-5} \pm 5.78\text{e}{-5}$	151.0
	C2F-BVI	$2.16\text{e}{-4} \pm 4.38\text{e}{-4}$	**140.8**

Table 3: Accuracy of lifted methods for RKFs against GaBP.

a normal distribution, i.e., $w \sim \mathcal{N}(0, Q)$, and $v \sim \mathcal{N}(0, R)$, for covariance matrices Q and R. The RKF model defines a lifted KF, i.e. similar state variables and similar observation nodes share the same transition and observation model.

We use extracted groundwater level data from the Republican River Compact Association model [McKusick, 2003]. The data set contains a record over 850 months of the water level data for 3,420 wells. We followed the same data preprocessing steps as in [2015], where wells in the same area are grouped together and are assumed to share same transition and observation0 model. We test our algorithms on two different structure settings, a tree-structured model and a model with cycles. For the tree-structured model, we define the matrices $A = \alpha \cdot I$, $C = I$, $Q = \beta \cdot I$, $R = \gamma \cdot I$ where $\alpha \sim U(0.5, 1)$, $\beta \sim U(5, 10)$, $\gamma \sim U(1, 5)$ and I is the identity matrix. For general model, we select $A = \alpha \cdot I + 0.01$ and $Q = \beta \cdot J$, where J is the matrix of ones, and other matrices are as before. We chose 20 months of record as the observation of the model. Note that the model defined has a linear Gaussian potential $\exp((x_{t+1} - Ax_t)^2/\sigma^2)$ with the number of state variables as its dimension, which makes inference challenging. For simplicity, we expressed the model as a product of pairwise potentials.

We compared our VI methods and Particle BP (EPBP and HLBP) methods against GaBP. The VI methods used Adam optimizer with learning rate 0.2; BP used 20 particle points.

Table 3 reports the resulting average ℓ_1 difference and KL-divergence of the MAP estimate of the last time step nodes against GaBP. The VI methods are accurate in this model, owing mostly to the unimodality of the marginals, providing evidence in support of **Q1** for this setting, even when the graph contains cycles. The VI methods also obtained better KL-divergence than Particle BP methdos, and the KL-divergence of all methods would likely further reduce with additional mixture components/particles.

Significant speed-ups are obtained by the lifted methods with a small improvement in the case of C2F. The timing results indicate significant performance improvements from lifting, answering **Q2** affirmatively. However, as the implementations were not optimized for performance in this case, the timing comparison between BVI vs EPBP is misleading. As the VI methods can also be efficiently implemented in Tensorflow to exploit parallel hardware (note this is inherently difficult for a sampling method like EPBP), we also provide timing experiments in a high performance setting for BVI/NPVI and their lifted versions. For BVI/L-BVI, the TensorFlow implementation took $107.8/\mathbf{5.6}s$ to run on tree model and $16.3/\mathbf{12.6}s$ on cycle model; NPVI/L-NPVI took $105.7/\mathbf{5.11}s$ and $15.7/\mathbf{12.8}s$ on tree and cycle models respectively, and gave identical performance to BVI/L-BVI.

Finally, comparing to the most recent hybrid lifted algorithm, HLBP [Chen *et al.*, 2019], although the table shows that HLBP is faster, the timing results are again incomparable because of implementation differences between BP and VI methods (our Tensorflow implementation of lifted VI methods are much faster). In this case, HLBP is significantly less accurate than VI, likely due to noise introduced by sampling. We also observed that the convergence of HLBP is very sensitive to the parameters of the model.

5 Discussion

We presented distribution-independent, model-agnostic hybrid lifted inference algorithms that makes minimal assumptions on the underlying distribution, and a simple C2F approach for handling symmetry breaking as a result of continuous evidence. We showed experimentally that the lifted and the C2F VI methods compare favorably in terms of accuracy against exact and particle-based methods for MAP and marginal inference tasks and can yield speed-ups over their non-lifted counterparts that range from moderate to significant, depending on the amount of evidence and the distribution of the evidence values. Additionally, we proved a sufficient condition under which the BFE over the marginal polytope is bounded from below, yielding a nontrivial approximation to the partition function, which supports our variational approach and may be of independent theoretical interest.

Acknowledgements

This work was supported, in part, by the DARPA Explainable Artificial Intelligence (XAI) program (N66001-17-2-4032). Sriraam Natarajan acknowledges the support of NSF grant IIS-1836565 and AFOSR award FA9550-18-1- 0462. Any opinions, findings and conclusions or recommendations are those of the authors and do not necessarily reflect the view of the DARPA, the Air Force, or the US government.

A Proof of Theorem 1

Theorem. *If there exists a collection of densities $g_i \in \mathcal{P}(\mathcal{X}_i)$ for each $i \in V$ such that $\sup_{x\in\mathcal{X}} \frac{p(x)}{\prod_i g_i(x_i)} < \infty$, then*

$$\inf_{q\in\mathcal{P}(\mathcal{X})} -\mathbb{E}_q[\log p] - \mathbb{H}_{\mathrm{B}}[q] > -\infty,$$

where $\mathcal{P}(\mathcal{X})$ is the set of all probability densities over $\mathcal{X}$.

Proof. The BFE can be expressed as

$$\begin{aligned}\mathcal{F}_{\mathrm{B}}(q) &= \mathbb{E}_q[-\log p] - \sum_{i=1}^{n}\mathbb{H}[q_i] + \sum_{c\in\mathcal{C}}\mathbb{T}[q_c] \\ &\geq \mathbb{E}_q[-\log p] - \sum_{i=1}^{n}\mathbb{H}[q_i] \\ &\triangleq \mathcal{L}(q),\end{aligned}$$

where the inequality follows from the fact that clique total correlation,

$$\mathbb{T}[q_c] := KL(q_c||\prod_{i\in c} q_i) = \mathbb{E}_{q_c(x_c)}[\log\left(\frac{q_c(x_c)}{\prod_{i\in c} q_i(x_i)}\right)]$$

is non-negative.

We will show that, under the conditions of the theorem, $\inf_{q\in\mathcal{P}(x)} \mathcal{L}(q) > -\infty$. To see this, we can reformulate the minimization of $\mathcal{L}$ as a convex optimization problem.

$$\inf_{q,m_{i\in\mathcal{V}}} \mathbb{E}_q[-\log p] - \sum_{i=1}^{n}\mathbb{H}[m_i]$$

subject to

$$\begin{gathered} q \in \mathcal{P}(x) \\ \forall i \in \mathcal{V}, x_i \in \mathcal{X}_i, \int_{x_{-i}} q(x) = m_i(x_i)\end{gathered}$$

Using the method of Lagrange multipliers, we construct the following dual optimization problem.

$$\sup_{g_{i\in\mathcal{V}}: g_i\in\mathcal{P}(\mathcal{X}_i)} \inf_{x\in\mathcal{X}} \left[\log \frac{\prod_i g_i(x_i)}{p(x)}\right]$$

So, in order for $\mathcal{L}$ to be bounded from below, it suffices that there exists a collection of densities, $g_{i\in\mathcal{V}}$, such that $\sup_{x\in\mathcal{X}} \frac{p(x)}{\prod_i g_i(x_i)} < \infty$.

□

References

[Ahmadi *et al.*, 2011] Babak Ahmadi, Kristian Kersting, and Scott Sanner. Multi-evidence lifted message passing, with application to PageRank and the Kalman filter. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

[Bui *et al.*, 2014] Hung Hai Bui, Tuyen N Huynh, and David Sontag. Lifted tree-reweighted variational inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[Chen *et al.*, 2019] Yuqiao Chen, Nicholas Ruozzi, and Sriraam Natarajan. Lifted message passing for hybrid probabilistic inference. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[Choi and Amir, 2012] Jaesik Choi and Eyal Amir. Lifted relational variational inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[Choi *et al.*, 2010] Jaesik Choi, Eyal Amir, and David Hill. Lifted inference for relational continuous models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

[Choi *et al.*, 2011] Jaesik Choi, Abner Guzman-Rivera, and Eyal Amir. Lifted relational Kalman filtering. In *Conference on Artificial Intelligence (AAAI)*, 2011.

[Choi *et al.*, 2015] Jaesik Choi, Eyal Amir, Tianfang Xu, and Albert J. Valocchi. Learning relational Kalman filtering. In *Conference on Artificial Intelligence (AAAI)*, 2015.

[Cseke and Heskes, 2011] Botond Cseke and Tom Heskes. Properties of Bethe free energies and message passing in Gaussian models. *Journal of Artificial Intelligence Research*, pages 1–24, 2011.

[Gallo and Ihler, 2018] Nicholas Gallo and Alexander Ihler. Lifted generalized dual decomposition. In *Conference on Artificial Intelligence (AAAI)*, 2018.

[Gershman *et al.*, 2012] Samuel Gershman, Matt Hoffman, and David Blei. Nonparametric variational inference. In *International Conference on Machine Learning (ICML)*, pages 235–242, 2012.

[Golub and Welsch, 1969] Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of computation*, 23(106):221–230, 1969.

[Guo *et al.*, 2019a] Yuanzhen Guo, Hao Xiong, and Nicholas Ruozzi. Marginal inference in continuous Markov random fields using mixtures. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 33, pages 7834–7841, 2019.

[Guo *et al.*, 2019b] Yuanzhen Guo, Hao Xiong, Yibo Yang, and Nicholas Ruozzi. One-shot marginal MAP inference in Markov random fields. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[Habeeb *et al.*, 2017] Haroun Habeeb, Ankit Anand, Mausam, and Parag Singla. Coarse-to-fine lifted map inference in computer vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[Jaakkola and Jordan, 1998] Tommi S. Jaakkola and Michael Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pages 163–173. Springer, 1998.

[Kersting *et al.*, 2009] Kristian Kersting, Babak Ahmadi, and Sriraam Natarajan. Counting belief propagation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Lienart *et al.*, 2015] Thibaut Lienart, Yee Whye Teh, and Arnaud Doucet. Expectation particle belief propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[Liu and Wang, 2016] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[McKusick, 2003] Vincent L. McKusick. Final report for the special master with certificate of adoption of RRCA groundwater model. *U.S. Supreme Court*, 2003.

[Ruozzi, 2012] Nicholas Ruozzi. The Bethe partition function of log-supermodular graphical models. In *Neural Information Processing Systems (NeurIPS)*, 2012.

[Ruozzi, 2013] Nicholas Ruozzi. Beyond log-supermodularity: Lower bounds and the Bethe partition function. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[Ruozzi, 2017] Nicholas Ruozzi. A lower bound on the partition function of attractive graphical models in the continuous case. In *Artificial Intelligence and Statistics (AISTATS)*, 2017.

[Singla and Domingos, 2008] Parag Singla and Pedro Domingos. Lifted first-order belief propagation. In *Twenty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

[Wang and Domingos, 2008] Jue Wang and Pedro Domingos. Hybrid markov logic networks. In *Conference on Artificial Intelligence (AAAI)*, 2008.

[Welling and Teh, 2001] Max Welling and Yee Whye Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2001.

[Yedidia *et al.*, 2001] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Neural Information Processing Systems (NeurIPS)*, 2001.

[Yedidia *et al.*, 2005] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282 – 2312, July 2005.