Lifted Tree-Reweighted Variational Inference

Hung Hai Bui
Natural Language Understanding Lab
Nuance Communications
Sunnyvale, CA, USA
bui.h.hung@gmail.com

Tuyen N. Huynh
John von Neumann Institute
Vietnam National University
Ho Chi Minh City
tuyen.huynh@jvn.edu.vn

David Sontag
Courant Institute of Mathematical Sciences
New York University
dsontag@cs.nyu.edu

Abstract

We analyze variational inference for highly symmetric graphical models such as those arising from first-order probabilistic models. We first show that for these graphical models, the tree-reweighted variational objective lends itself to a compact lifted formulation which can be solved much more efficiently than the standard TRW formulation for the ground graphical model. Compared to earlier work on lifted belief propagation, our formulation leads to a convex optimization problem for lifted marginal inference and provides an upper bound on the partition function. We provide two approaches for improving the lifted TRW upper bound. The first is a method for efficiently computing maximum spanning trees in highly symmetric graphs, which can be used to optimize the TRW edge appearance probabilities. The second is a method for tightening the relaxation of the marginal polytope using lifted cycle inequalities and novel exchangeable cluster consistency constraints.

1 Introduction

Lifted probabilistic inference focuses on exploiting symmetries in probabilistic models for efficient inference [5, 2, 3, 10, 17, 18, 21]. Work in this area has demonstrated the possibility to perform very efficient inference in highly-connected, large tree-width, but *symmetric* models, such as those arising in the context of relational (first-order) probabilistic models and exponential family random graphs [19]. These models also arise frequently in probabilistic programming languages, an area of increasing importance as demonstrated by DARPA's PPAML program (Probabilistic Programming for Advancing Machine Learning).

Even though lifted inference can sometimes offer order-of-magnitude improvement in performance, approximation is still necessary. A topic of particular interest is the interplay between lifted inference and variational approximate inference. Lifted loopy belief propagation (LBP) [13, 21] was one of the first attempts at exploiting symmetry to speed up loopy belief propagation; subsequently, counting belief propagation (CBP) [16] provided additional insights into the nature of symmetry in BP. Nevertheless, these work were largely procedural and specific to the choice of message-passing algorithm (in this case, loopy BP). More recently, Bui et al., [3] proposed a general framework for lifting a broad class of convex variational techniques by formalizing the notion of symmetry (defined via automorphism groups) of graphical models and the corresponding variational optimization problems themselves, independent of any specific methods or solvers.

Our goal in this paper is to extend the lifted variational framework in [3] to address the important case of approximate marginal inference. In particular, we show how to lift the tree-reweighted (TRW) convex formulation of marginal inference [28]. As far as we know, our work presents the first lifted *convex* variational marginal inference, with the following benefits over previous work: (1) a lifted convex upper bound of the log-partition function, (2) a new tightening of the relaxation of the lifted marginal polytope exploiting exchangeability, and (3) a convergent inference algorithm. We note that convex upper bounds of the log-partition function immediately lead to concave lower bounds of the log-likelihood which can serve as useful surrogate loss functions in learning and parameter estimation [29, 13].

To achieve the above goal, we first analyze the symmetry of the TRW log-partition and entropy bounds. Since TRW bounds depend on the choice of the edge appearance probabilities ρ, we prove that the quality of the TRW bound is not affected if one only works with suitably symmetric ρ. Working with symmetric ρ gives rise to an explicit lifted formulation of the TRW optimization problem that is equivalent but much more compact. This convex objective function can be convergently optimized via a Frank-Wolfe (conditional gradient) method where each Frank-Wolfe iteration solves a lifted MAP inference problem. We then discuss the optimization of the edge-appearance vector ρ, effectively yielding a lifted algorithm for computing maximum spanning trees in symmetric graphs.

As in Bui et al.'s framework, our work can benefit from any tightening of the local polytope such as the use of cycle inequalities [1, 23]. In fact, each method for relaxing the marginal polytope immediately yields a variant of our algorithm. Notably, in the case of exchangeable random variables, radically sharper tightening (sometimes even exact characterization of the lifted marginal polytope) can be obtained via a set of simple and elegant linear constraints which we call *exchangeable polytope constraints*. We provide extensive simulation studies comparing the behaviors of different variants of our algorithm with exact inference (when available) and lifted LBP demonstrating the advantages of our approach. The supplementary material [4] provides additional proof and algorithm details.

2 Background

We begin by reviewing variational inference and the tree-reweighted (TRW) approximation. We focus on inference in Markov random fields, which are distributions in the exponential family given by $\Pr(x;\theta) = \exp\{\langle\Phi(x),\theta\rangle - A(\theta)\}$, where $A(\theta)$ is called the *log-partition function* and serves to normalize the distribution. We assume that the random variables $x \in \mathcal{X}^n$ are discrete-valued, and that the features $(\Phi_i),\ i \in \mathcal{I}$ factor according to the graphical model structure $\mathcal{G}$; Φ can be non-pairwise and is assumed to be overcomplete. This paper focuses on the inference tasks of estimating the marginal probabilities $p(x_i)$ and approximating the log-partition function. Throughout the paper, the domain $\mathcal{X}$ is the binary domain $\{0,1\}$, however, except for the construction of exchangeable polytope constraints in Section 6, this restriction is not essential.

Variational inference approaches view the log-partition function as a convex optimization problem over the marginal polytope $A(\theta) = \sup_{\mu\in\mathcal{M}(\mathcal{G})}\langle\mu,\theta\rangle - A^*(\mu)$ and seek tractable approximations of the negative entropy A^* and the marginal polytope $\mathcal{M}$ [27]. Formally, $-A^*(\mu)$ is the entropy of the maximum entropy distribution with moments μ. Observe that $-A^*(\mu)$ is upper bounded by the entropy of the maximum entropy distribution consistent with any subset of the expected sufficient statistics μ. To arrive at the TRW approximation [26], one uses a subset given by the pairwise moments of a spanning tree[1]. Hence for any distribution ρ over spanning trees, an upper bound on $-A^*$ is obtained by taking a convex combination of tree entropies $-B^*(\tau,\rho) = \sum_{s\in V(G)} H(\tau_s) - \sum_{e\in E(G)} I(\tau_e)\rho_e$. Since ρ is a distribution over spanning trees, it must belong to the spanning tree polytope $\mathbb{T}(\mathcal{G})$ with ρ_e denoting the edge appearance probability of e. Combined with a relaxation of the marginal polytope OUTER $\supset \mathcal{M}$, an upper bound B of the log-partition function is obtained:

$$A(\theta) \le B(\theta,\rho) = \sup_{\tau\in\text{OUTER}(\mathcal{G})} \langle\tau,\theta\rangle - B^*(\tau,\rho) \qquad (1)$$

We note that B^* is linear w.r.t. ρ, and for $\rho \in \mathbb{T}(G)$, B^* is convex w.r.t. τ. On the other hand, B is convex w.r.t. ρ and θ.

The optimal solution $\tau^*(\rho,\theta)$ of the optimization problem (1) can be used as an approximation to the mean parameter $\mu(\theta)$. Typically, the local polytope LOCAL given by pairwise consistency constraints is used as the relaxation OUTER; in this paper, we also consider tightening of the local polytope.

Since (1) holds with any edge appearance ρ in the spanning tree polytope $\mathbb{T}$, the TRW bound can be further improved by optimizing ρ

$$\inf_{\rho\in\mathbb{T}(G)} B(\theta,\rho) \qquad (2)$$

The resulting ρ^* is then plugged into (1) to find the marginal approximation. In practice, one might choose to work with some fixed choice of ρ, for example the uniform distribution over all spanning trees. [14] proposed using the most uniform edge-weight $\arg\inf_{\rho\in\mathbb{T}(G)} \sum_{e\in E}(\rho_e - \frac{|V|-1}{|E|})^2$ which can be found via conditional gradient where each direction-finding step solves a maximum spanning tree problem.

Several algorithms have been proposed for optimizing the TRW objective (1) given fixed edge appearance probabilities. [27] derived the tree-reweighted belief propagation algorithm from the fixed point conditions. [8] show how to solve the dual of the TRW objective, which is a geometric program. Although this algorithm has the advantage of guaranteed convergence, it is non-trivial to generalize this approach to use tighter relaxations of the marginal polytope, which we show is essential for lifted inference. [14] use an explicit set of spanning trees and then use dual decomposition to solve the optimization problem. However, as we show in the next section, to maintain symmetry it is essential that one *not* work directly with spanning trees but rather use symmetric edge appearance probabilities. [23] optimize TRW over the local and cycle polytopes using a Frank-Wolfe (conditional gradient) method, where each iteration requires solving a linear program. We follow this latter approach in our paper.

To optimize the edge appearance in (2), [26] proposed using conditional gradient. They observed that $\frac{\partial B(\theta,\rho)}{\partial\rho_e} = -\frac{\partial B^*(\tau^*,\rho)}{\partial\rho_e} = -I(\tau_e^*)$ where τ^* is the solution of (1). The direction-finding step in conditional gradient reduces to solving $\sup_{\rho\in\mathbb{T}}\langle\rho, I\rangle$, again equivalent to finding the maximum spanning tree with edge mutual information $I(\tau_e^*)$ as weights. We discuss the corresponding lifted problem in section 5.

[1] If the original model contains non-pairwise potentials, they can be represented as cliques in the graphical model, and the bound based on spanning trees still holds.

3 Lifted Variational Framework

We build on the key element of the lifted variational framework introduced in [3]. The automorphism group of a graphical model, or more generally, an exponential family is defined as the group $\mathbb{A}$ of permutation pairs (π, γ) where π permutes the set of variables and γ permutes the set of features in such a way that they preserve the feature function: $\Phi^{\gamma^{-1}}(x^{\pi}) = \Phi(x)$. Note that this construction of $\mathbb{A}$ is entirely based on the structure of the model and does not depend on the particular choice of the model parameters; nevertheless the group stabilizes[2] (preserves) the key characteristics of the exponential family such as the marginal polytope $\mathcal{M}$, the log-partition A and entropy $-A^*$.

As shown in [3] the automorphism group is particularly useful for exploiting symmetries when parameters are tied. For a given parameter-tying partition Δ such that $\theta_i = \theta_j$ for i, j in the same cell[3] of Δ, the group $\mathbb{A}$ gives rise to a subgroup called the lifting group $\mathbb{A}_\Delta$ that stabilizes the tied-parameter vector θ as well as the exponential family. The orbit partition of the the lifting group can be used to formulate equivalent but more compact variational problems. More specifically, let $\varphi = \varphi(\Delta)$ be the orbit partition induced by the lifting group on the feature index set $\mathcal{I} = \{1 \dots m\}$, let $\mathbb{R}^m_{[\varphi]}$ denote the symmetrized subspace $\{r \in \mathbb{R}^m \text{ s.t. } r_i = r_j \ \forall i, j \text{ in the same cell of } \varphi\}$ and define the lifted marginal polytope $\mathcal{M}_{[\varphi]}$ as $\mathcal{M} \cap \mathbb{R}^m_{[\varphi]}$, then (see Theorem 4 of [3])

$$\sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle - A^*(\mu) = \sup_{\mu \in \mathcal{M}_{[\varphi]}} \langle \theta, \mu \rangle - A^*(\mu) \quad (3)$$

In practice, we need to work with convex variational approximations of the LHS of (3) where $\mathcal{M}$ is relaxed to an outer bound OUTER($\mathcal{G}$) and A^* is approximated by a convex function $B^*(\mu)$. We now state a similar result for lifting general convex approximations.

Theorem 1. *If $B^*(\mu)$ is convex and stabilized by the lifting group $\mathbb{A}_\Delta$, i.e., for all $(\pi, \gamma) \in \mathbb{A}_\Delta$, $B^*(\mu^\gamma) = B^*(\mu)$, then φ is the lifting partition for the approximate variational problem*

$$\sup_{\mu \in \text{OUTER}(\mathcal{G})} \langle \theta, \mu \rangle - B^*(\mu) = \sup_{\mu \in \text{OUTER}_{[\varphi]}} \langle \theta, \mu \rangle - B^*(\mu) \quad (4)$$

The importance of Theorem 1 is that it shows that it is equivalent to optimize over a subset of OUTER($\mathcal{G}$) where pseudo-marginals in the same orbit are restricted to take the same value. As we will show in Section 4.2, this will allow us to combine many of the terms in the objective, which is where the computational gains will derive from. A sketch of its proof is as follows. Consider a single pseudo-marginal vector μ. Since the objective value is the same for every μ^γ for $(\pi, \gamma) \in \mathbb{A}_\Delta$ and since the objective is concave, the *average* of these, $\frac{1}{|\mathbb{A}_\Delta|} \sum_{(\pi,\gamma) \in \mathbb{A}_\Delta} \mu^\gamma$, must have at least as good of an objective value. Furthermore, note that this averaged vector lives in the symmetrized subspace. Thus, it suffices to optimize over $\text{OUTER}_{[\varphi]}$.

4 Lifted Tree-Reweighted Problem

4.1 Symmetry of TRW Bounds

We now show that Theorem 1 can be used to lift the TRW optimization problem (1). Note that the applicability of Theorem 1 is not immediately obvious since B^* depends on the distribution over trees implicit in ρ. In establishing that the condition in Theorem 1 holds, we need to be careful so that the choice of the distribution over trees ρ does not destroy the symmetry of the problem.

The result below ensures that with no loss in optimality, ρ can be assumed to be suitably symmetric. More specifically, let $\varphi^E = \varphi^E(\Delta)$ be the set of $\mathcal{G}$'s edge orbits induced by the action of the lifting group $\mathbb{A}_\Delta$; the edge-weights ρ_e for every e in the same edge orbits can be constrained to be the same, i.e. ρ can be restricted to $\mathbb{T}_{[\varphi^E]}$.

Theorem 2. *For any $\rho \in \mathbb{T}$, there exists a symmetrized $\hat{\rho} \in \mathbb{T}_{[\varphi^E]}$ that yields at least as good an upper bound, i.e.*

$$B(\theta, \hat{\rho}) \leq B(\theta, \rho) \ \ \forall \theta \in \Theta_{[\Delta]}$$

As a consequence, in optimizing the edge appearance, ρ can be restricted to the symmetrized spanning tree polytope $\mathbb{T}_{[\varphi^E]}$

$$\forall \theta \in \Theta_{[\Delta]}, \ \inf_{\rho \in \mathbb{T}} B(\theta, \rho) = \inf_{\rho \in \mathbb{T}_{[\varphi^E]}} B(\theta, \rho)$$

Proof. Let ρ be the argmin of the LHS, and define $\hat{\rho} = \frac{1}{|\mathbb{A}_\Delta|} \sum_{\pi \in \mathbb{A}_\Delta} \rho^\pi$ so that $\hat{\rho} \in \mathbb{T}_{[\varphi^E]}$. For all $(\pi, \gamma) \in \mathbb{A}_\Delta$ and for all tied-parameter $\theta \in \Theta_{[\Delta]}$, $\theta^\pi = \theta$, so $B(\theta, \rho^\pi) = B(\theta^\pi, \rho^\pi)$. By theorem 1 of [3], π must be an automorphism of the graph $\mathcal{G}$. By lemma 7 (see supplementary material), $B(\theta^\pi, \rho^\pi) = B(\theta, \rho)$. Thus $B(\theta, \rho^\pi) = B(\theta, \rho)$. Since B is convex w.r.t. ρ, by Jensen's inequality we have that $B(\theta, \hat{\rho}) \leq \frac{1}{|\mathbb{A}_\Delta|} \sum_{\pi \in \mathbb{A}_\Delta} B(\theta, \rho^\pi) = B(\theta, \rho)$. □

Using a symmetric choice of ρ, the TRW bound B^* then satisfies the condition of theorem 1, enabling the applicability of the general lifted variational inference framework.

Theorem 3. *For a fixed $\rho \in \mathbb{T}_{[\varphi^E]}$, φ is the lifting partition for the TRW variational problem*

$$\sup_{\tau \in \mathit{OUTER}(\mathcal{G})} \langle \tau, \theta \rangle - B^*(\tau, \rho) = \sup_{\tau \in \mathit{OUTER}_{[\varphi]}} \langle \tau, \theta \rangle - B^*(\tau, \rho) \quad (5)$$

[2] Formally, $\mathbb{G}$ stabilizes x if $x^g = x$ for all $g \in \mathbb{G}$.

[3] If $\Delta = \{\Delta_1 \dots \Delta_K\}$ is a partition of S, then each subset $\Delta_k \subset S$ is called a cell.

4.2 Lifted TRW Problems

We give the explicit lifted formulation of the TRW optimization problem (5). As in [3], we restrict τ to OUTER$_{[\varphi]}$ by introducing the lifted variables $\bar{\tau}_j$ for each cell φ_j, and for all $i \in \varphi_j$, enforcing that $\tau_i = \bar{\tau}_j$. Effectively, we substitute every occurrence of τ_i, $i \in \varphi_j$ by $\bar{\tau}_j$; in vector form, τ is substituted by $D\bar{\tau}$ where D is the characteristic matrix of the partition φ: $D_{ij} = 1$ if $i \in \varphi_j$ and 0 otherwise. This results in the lifted form of the TRW problem

$$\sup_{D\bar{\tau} \in \text{OUTER}} \left\langle \bar{\tau}, \bar{\theta} \right\rangle - \overline{B^*}(\bar{\tau}, \bar{\rho}) \tag{6}$$

where $\bar{\theta} = D^\top \theta$; $\overline{B^*}$ is obtained from B^* via the above substitution; and $\bar{\rho}$ is the edge appearance per edge-orbit: for every edge orbit $\mathbf{e}$, and for every edge $e \in \mathbf{e}$, $\rho_e = \bar{\rho}_\mathbf{e}$. Using an alternative but equivalent form $B^* = -\sum_{v \in V}(1 - \sum_{e \in Nb(v)} \rho_e) H(\tau_v) - \sum_{e \in E} \rho_e H(\tau_e)$, we obtain the following explicit form for

$$\begin{aligned} \overline{B^*}(\bar{\tau}, \bar{\rho}) &= -\sum_{\mathbf{v} \in \bar{V}} \left(|\mathbf{v}| - \sum_{\mathbf{e} \in N(\mathbf{v})} |\mathbf{e}| d(\mathbf{v}, \mathbf{e}) \bar{\rho}_\mathbf{e} \right) H(\bar{\tau}_\mathbf{v}) \\ &\quad - \sum_{\mathbf{e} \in \bar{E}} |\mathbf{e}| \bar{\rho}_\mathbf{e} H(\bar{\tau}_\mathbf{e}) \end{aligned} \tag{7}$$

Intuitively, the above can be viewed as a combination of node and edge entropies defined on nodes and edges of the lifted graph $\bar{\mathcal{G}}$. Nodes of $\bar{\mathcal{G}}$ are the node orbits of $\mathcal{G}$ while edges are the edge-orbits of $\mathcal{G}$. $\bar{\mathcal{G}}$ is not a simple graph: it can have self-loops or multi-edges between the same node pair (see Fig. 1). We encode the incidence on this graph as follows: $d(\mathbf{v}, \mathbf{e}) = 0$ if $\mathbf{v}$ is not incident to $\mathbf{e}$, $d(\mathbf{v}, \mathbf{e}) = 1$ if $\mathbf{v}$ is incident to $\mathbf{e}$ and $\mathbf{e}$ is not a loop, $d(\mathbf{v}, \mathbf{e}) = 2$ if $\mathbf{e}$ is a loop incident to $\mathbf{v}$. The *entropy* at the node orbit $\mathbf{v}$ is defined as

$$H(\bar{\tau}_\mathbf{v}) = -\sum_{t \in \mathcal{X}} \bar{\tau}_{\mathbf{v}:t} \ln(\bar{\tau}_{\mathbf{v}:t})$$

and the entropy at the edge orbit $\mathbf{e}$ is

$$H(\bar{\tau}_\mathbf{e}) = -\sum_{t,h \in \mathcal{X}} \bar{\tau}_{\overline{\{e_1:t, e_2:h\}}} \ln(\bar{\tau}_{\overline{\{e_1:t, e_2:h\}}})$$

where $\{e_1, e_2\}$ for $e_1, e_2 \in V$ is a representative (any element) of $\mathbf{e}$, $\{e_1{:}t, e_2{:}h\}$ is an assignment of the ground edge $\{e_1, e_2\}$, and $\overline{\{e_1{:}t, e_2{:}h\}}$ is the assignment orbit. As in [3], we write $\overline{\{e_1{:}t, e_2{:}t\}}$ as $\mathbf{e}{:}t$, and for $t < h$, $\overline{\{e_1{:}t, e_2{:}h\}}$ as $\mathbf{a}{:}(t, h)$ where $\mathbf{a}$ is the arc-orbit $\overline{(e_1, e_2)}$.

When OUTER is the local or cycle polytope, the constraints $D\bar{\tau} \in$ OUTER yield the lifted local (or cycle) polytope respectively. For these constraints, we use the same form given in [3]. In section 6, we describe a set of additional constraints for further tightening when some cluster of nodes are exchangeable.

Example. Consider the MRF shown in Fig. 1 (left) with 10 binary variables that we denote B_i (for the blue nodes) and R_i (for the red nodes). The node and edge coloring denotes shared parameters. Let θ_b and θ_r be the single-node potentials used for the blue and red nodes, respectively. Let θ_{r_e} be the edge potential used for the red edges connecting the blue and red nodes, θ_{b_e} for the edge potential used for the blue edges (B_i, B_{i+1}), and θ_{k_e} for the edge potential used for the black edges (B_i, B_{i+2}).

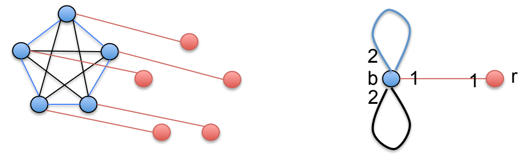

Figure 1: Left: ground graphical model. Same colored nodes and edges have the same parameters. Right: lifted graph showing 2 node orbits ($\mathbf{b}$ and $\mathbf{r}$), and 3 edge orbits. Numbers on the lifted graph representing the incidence degree $d(\mathbf{v}, \mathbf{e})$ between an edge and a node orbit.

There are two node orbits: $\mathbf{b} = \{B_1, \ldots, B_5\}$ and $\mathbf{r} = \{R_1, \ldots, R_5\}$. There are three edge orbits: $\mathbf{r_e}$ for the red edges, $\mathbf{b_e}$ for the blue edges , and $\mathbf{k_e}$ for the black edges. The size of the node and edge orbits are all 5 (e.g., $|\mathbf{b}| = |\mathbf{b_e}| = 5$), and $d(\mathbf{b}, \mathbf{r_e}) = d(\mathbf{r}, \mathbf{r_e}) = 1$, $d(\mathbf{b}, \mathbf{b_e}) = d(\mathbf{b}, \mathbf{k_e}) = 2$. Suppose that ρ corresponds to a uniform distribution over spanning trees, which satisfies the symmetry needed by Theorem 2. We then have $\bar{\rho}_{\mathbf{r_e}} = 1$ and $\bar{\rho}_{\mathbf{b_e}} = \bar{\rho}_{\mathbf{k_e}} = \frac{2}{5}$. Putting all of this together, the lifted TRW entropy is given by $\overline{B^*}(\bar{\tau}, \bar{\rho}) = 8H(\bar{\tau}_\mathbf{b}) - 5H(\bar{\tau}_{\mathbf{r_e}}) - 2H(\bar{\tau}_{\mathbf{b_e}}) - 2H(\bar{\tau}_{\mathbf{k_e}})$. We illustrate the expansion of the entropy of the red edge orbit $H(\bar{\tau}_{\mathbf{r_e}})$. This edge orbit has 2 corresponding arc-orbits: $\mathbf{rb_a} = \{(R_i, B_i)\}$ and $\mathbf{br_a} = \{(B_i, R_i)\}$. Thus, $H(\bar{\tau}_{\mathbf{r_e}}) = -\bar{\tau}_{\mathbf{r_e}:00} \ln \bar{\tau}_{\mathbf{r_e}:00} - \bar{\tau}_{\mathbf{r_e}:11} \ln \bar{\tau}_{\mathbf{r_e}:11} - \bar{\tau}_{\mathbf{rb_a}:01} \ln \bar{\tau}_{\mathbf{rb_a}:01} - \bar{\tau}_{\mathbf{br_a}:01} \ln \bar{\tau}_{\mathbf{br_a}:01}$.

Finally, the linear term in the objective is given by $\left\langle \bar{\tau}, \bar{\theta} \right\rangle = 5\left\langle \bar{\tau}_\mathbf{b}, \theta_b \right\rangle + 5\left\langle \bar{\tau}_\mathbf{r}, \theta_r \right\rangle + 5\left\langle \bar{\tau}_{\mathbf{r_e}}, \theta_{r_e} \right\rangle + 5\left\langle \bar{\tau}_{\mathbf{b_e}}, \theta_{b_e} \right\rangle + 5\left\langle \bar{\tau}_{\mathbf{k_e}}, \theta_{k_e} \right\rangle$ where, as an example, $\left\langle \bar{\tau}_{\mathbf{r_e}}, \theta_{r_e} \right\rangle = \bar{\tau}_{\mathbf{r_e}:00}\theta_{r_e,00} + \bar{\tau}_{\mathbf{r_e}:11}\theta_{r_e,11} + \bar{\tau}_{\mathbf{br_a}:01}\theta_{r_e,01} + \bar{\tau}_{\mathbf{rb_a}:01}\theta_{r_e,10}$

4.3 Optimization using Frank-Wolfe

What remains is to describe how to optimize Eq. 6. Our lifted tree-reweighted algorithm is based on Frank-Wolfe, also known as the conditional gradient method [7, 11]. First, we initialize with a pseudo-marginal vector corresponding to the uniform distribution, which is guaranteed to be in the lifted marginal polytope. Next, we solve the linear program whose objective is given by the gradient of the objective Eq. 6 evaluated at the current point, and whose constraint set is OUTER. When using the lifted cycle relaxation, we solve this linear program using a cutting-plane algorithm [3, 23]. We then perform a line search to find the optimal step size using a golden section search (a type of binary search that finds the maxima of a unimodal function), and finally repeat using the new pseudo-marginal vector. We warm start each linear program using the optimal basis found in the previous run, which makes the LP solves extremely fast after the first couple of iterations. Although we use a generic LP solver in our experiments, it is also possible to use dual decomposition to derive efficient algorithms specialized to graphical models [24].

5 Lifted Maximum Spanning Tree

Optimizing the TRW edge appearance probability ρ requires finding the maximum spanning tree (MST) in the ground graphical model. For lifted TRW, we need to perform MST while using only information from the node and edge orbits, without referring to the ground graph. In this section, we present a lifted MST algorithm for symmetric graphs which works at the orbit level.

Suppose that we are given a *weighted* graph $(\mathcal{G}, w)$, its automorphism group $\mathbb{A} = Aut(\mathcal{G})$ and its node and edge orbits. We aim to derive an algorithm analogous to the Kruskal's algorithm, but with complexity depends only on the number of node/edge orbits of $\mathcal{G}$. However, if the algorithm has to return an actual spanning tree of $\mathcal{G}$ then clearly its complexity cannot be less than $O(|V|)$. Instead, we consider an equivalent problem: solving a linear program on the spanning-tree polytope

$$\sup_{\rho \in \mathbb{T}(\mathcal{G})} \langle \rho, w \rangle \tag{8}$$

The same mechanism for lifting convex optimization problem (Lemma 1 in [3]) applies to this problem. Let φ^E be the edge orbit partition, then an equivalent lifted problem problem is

$$\sup_{\rho \in \mathbb{T}_{[\varphi^E]}} \langle \rho, w \rangle \tag{9}$$

Since ρ_e is constrained to be the same for edges in the same orbit, it is now possible to solve (9) with complexity depending only on the number of orbits. Any solution ρ of the LP (8) can be turned into a solution $\bar{\rho}$ of (9) by letting $\bar{\rho}(\mathbf{e}) = \frac{1}{|\mathbf{e}|} \sum_{e' \in \mathbf{e}} \rho(e')$.

5.1 Lifted Kruskal's Algorithm

The Kruskal's algorithm first sorts the edges according to their decreasing weight. Then starting from an empty graph, at each step it greedily attempts to add the next edge while maintaining the property that the used edges form a forest (containing no cycle). The forest obtained at the end of this algorithm is a maximum-weight spanning tree.

Imagine how Kruskal's algorithm would operate on a weighted graph $\mathcal{G}$ with non-trivial automorphisms. Let $\mathbf{e}_1, \dots, \mathbf{e}_k$ be the list of edge-orbits sorted in the order of decreasing weight (the weights w on all edges in the same orbit by definition must be the same). The main question therefore is how many edges in each edge-orbit $\mathbf{e}_i$ will be added to the spanning tree by the Kruskal's algorithm. Let $\mathcal{G}_i$ be the subgraph of $\mathcal{G}$ formed by the set of all the edges and nodes in $\mathbf{e}_1, \dots \mathbf{e}_i$. Let $V(\mathcal{G})$ and $C(\mathcal{G})$ denote the set of nodes and set of connected components of a graph, respectively. Then (see the supplementary material for proof)

Lemma 4. *The number of edges in $\mathbf{e}_i$ appearing in the MST found by the Kruskal's algorithm is $\delta_V^{(i)} - \delta_C^{(i)}$ where $\delta_V^{(i)} = |V(\mathcal{G}_i)| - |V(\mathcal{G}_{i-1})|$ and $\delta_C^{(i)} = |C(\mathcal{G}_i)| - |C(\mathcal{G}_{i-i})|$. Thus a solution for the linear program (9) is $\bar{\rho}(\mathbf{e}_i) = \frac{\delta_V^{(i)} - \delta_C^{(i)}}{|\mathbf{e}_i|}$.*

5.2 Lifted Counting of the Number of Connected Components

We note that counting the number of nodes can be done simply by adding the size of each node orbit. The remaining difficulty is how to count the number of connected components of a given graph[4] $\mathcal{G}$ using only information at the orbit level. Let $\bar{\mathcal{G}}$ be the lifted graph of $\mathcal{G}$. Then (see supplementary material for proof)

Lemma 5. *If $\bar{\mathcal{G}}$ is connected then all connected components of $\mathcal{G}$ are isomorphic. Thus if furthermore $\mathcal{G}'$ is a connected component of $\mathcal{G}$ then $|C(\mathcal{G})| = |V(\mathcal{G})|/|V(\mathcal{G}')|$.*

To find just one connected component, we can choose an arbitrary node u and compute $\bar{\mathcal{G}}[u]$, the lifted graph fixing u (see section 8.1 in [3]), then search for the connected component in $\bar{\mathcal{G}}[u]$ that contains $\{u\}$. Finally, if $\bar{\mathcal{G}}$ is not connected, we simply apply lemma 5 for each connected component of $\bar{\mathcal{G}}$.

The final lifted Kruskal's algorithm combines lemma 4 and 5 while keeping track of the set of connected components of $\bar{\mathcal{G}}_i$ incrementally. The full algorithm is given in the supplementary material.

6 Tightening via Exchangeable Polytope Constraints

One type of symmetry often found in first-order probabilistic models are large sets of exchangeable random variables. In certain cases, exact inference with exchangeable variables is possible via lifted counting elimination and its generalization [17, 2]. The drawback of these exact methods is that they do not apply to many models (e.g., those with transitive clauses). Lifted variational inference methods do not have this drawback, however local and cycle relaxation can be shown to be loose in the exchangeable setting, a potentially serious limitation compared to earlier work.

To remedy this situation, we now show how to take advantage of highly symmetric subset of variables to tighten the relaxation of the lifted marginal polytope.

We call a set of random variables χ an *exchangeable* cluster iff χ can be arbitrary permuted while preserving the probability model. Mathematically, the lifting group $\mathbb{A}_\Delta$ acts on χ and the image of the action is isomorphic to $\mathbb{S}(\chi)$,

[4] Since we are only interested in connectivity in this subsection, the weights of $\mathcal{G}$ play no role. Thus, orbits in this subsection can also be generated by the automorphism group of the unweighted version of $\mathcal{G}$.

the symmetric group on χ. The distribution of the random variables in χ is also exchangeable in the usual sense.

Our method for tightening the relaxation of the marginal polytope is based on lift-and-project, wherein we introduce auxiliary variables specifying the joint distribution of a large cluster of variables, and then enforce consistency between the cluster distribution and the pseudo-marginal vector [20, 24, 27]. In the ground model, one typically works with small clusters (e.g., triplets) because the number of variables grows exponentially with cluster size. The key (and nice) difference in the lifted case is that we can make use of very large clusters of highly symmetric variables: while the grounded relaxation would clearly blow up, the corresponding lifted relaxation can still remain compact.

Specifically, for an exchangeable cluster χ of arbitrary size, one can add cluster consistency constraints for the entire cluster and still maintain tractability. To keep the exposition simple, we assume that the variables are binary. Let $\mathfrak{C}$ denote a χ-configuration, i.e., a function $\mathfrak{C} : \chi \to \{0,1\}$. The set $\{\tau^{\chi}_{\mathfrak{C}} \mid \forall \text{ configuration } \mathfrak{C}\}$ is the collection of χ-cluster auxiliary variables. Since χ is exchangeable, all nodes in χ belong to the same node orbit; we call this node orbit $\mathbf{v}(\chi)$. Similarly, $\mathbf{e}(\chi)$ and $\mathbf{a}(\chi)$ denote the single edge and arc orbit that contains all edges and arcs in χ respectively. Let u_1, u_2 be two distinct nodes in χ. To enforce consistency between the cluster χ and the edge $\{u_1, u_2\}$ in the ground model, we introduce the constraints

$$\exists \tau^{\chi} : \sum_{\mathfrak{C}\ s.t.\ \mathfrak{C}(u_i)=s_i} \tau^{\chi}_{\mathfrak{C}} = \tau_{u_1:s_1,u_2:s_2} \quad \forall s_i \in \{0,1\} \quad (10)$$

These constraints correspond to using intersection sets of size two, which can be shown to be the exact characterization of the marginal polytope involving variables in χ if the graphical model only has pairwise potentials. If higher-order potentials are present, a tighter relaxation could be obtained by using larger intersection sets together with the techniques described below.

The constraints in (10) can be methodically lifted by replacing occurrences of ground variables with lifted variables at the orbit level. First observe that in place of the grounded variables $\tau_{u_1:s_1,u_2:s_2}$, the lifted local relaxation has three corresponding lifted variables, $\bar{\tau}_{\mathbf{e}(\chi):00}$, $\bar{\tau}_{\mathbf{e}(\chi):11}$ and $\bar{\tau}_{\mathbf{a}(\chi):01}$. Second, we consider the orbits of the set of configurations $\mathfrak{C}$. Since χ is exchangeable, there can be only $|\chi|+1$ χ-configuration orbits; each orbit contains all configurations with precisely k 1's where $k = 0 \ldots |\chi|$. Thus, instead of the $2^{|\chi|}$ ground auxiliary variables, we only need $|\chi|+1$ lifted cluster variables. Further manipulation leads to the following set of constraints, which we call *lifted exchangeable polytope* constraints.

Theorem 6. *Let χ be an exchangeable cluster of size n; $\mathbf{e}(\chi)$ and $\mathbf{a}(\chi)$ be the single edge and arc orbit of the graphical model that contains all edges and arcs in χ respectively; $\bar{\tau}$ be the lifted marginals. Then there exist $c^{\chi}_k \geq 0$, $k = 0 \ldots n$ such that*

$$\begin{aligned} \sum_{k=0}^{n-2} \frac{(n-k)(n-k-1)}{n(n-1)} c^{\chi}_k &= \bar{\tau}_{\mathbf{e}(\chi):00} \\ \sum_{k=0}^{n-2} \frac{(k+1)(k+2)}{n(n-1)} c^{\chi}_{k+2} &= \bar{\tau}_{\mathbf{e}(\chi):11} \\ \sum_{k=0}^{n-2} \frac{(n-k-1)(k+1)}{n(n-1)} c^{\chi}_{k+1} &= \bar{\tau}_{\mathbf{a}(\chi):01} \end{aligned}$$

Proof. See the supplementary material. □

In contrast to the lifted local and cycle relaxations, the number of variables and constraints in the lifted exchangeable relaxation depends linearly on the domain size of the first-order model. From the lifted local constraints given by [3], $\bar{\tau}_{\mathbf{e}(\chi):00} + \bar{\tau}_{\mathbf{e}(\chi):11} + 2\bar{\tau}_{\mathbf{a}(\chi):01} = 1$. Substituting in the expression involved $\tilde{c}^{\chi}_k$, we get $\sum_{k=0}^{n} c^{\chi}_k = 1$. Intuitively, c^{χ}_k represents the approximation of the marginal probability $\Pr(\sum_{i\in\chi} x_i = k)$ of having precisely k ones in χ.

As proved by [2], groundings of unary predicates in Markov Logic Networks (MLNs) gives rise to exchangeable clusters. Thus, for MLNs, the above theorem immediately suggests a tightening of the relaxation: for every unary predicate of a MLN, add a new set of constraints as above to the existing lifted local (or cycle) optimization problem. Although it is not the focus of our paper, we note that this should also improve the lifted MAP inference results of [3]. For example, in the case of a symmetric complete graphical model, lifted MAP inference using the linear program given by these new constraints would find the exact k that maximizes $\Pr(x_{\chi})$, hence recover the same solution as counting elimination. Marginal inference may still be inexact due to the tree-reweighted entropy approximation. We re-emphasize that the complexity of variational inference with lifted exchangeable constraints is guaranteed to be polynomial in the domain size, unlike exact methods based on lifted counting elimination and variable elimination.

7 Experiments

In this section, we provide an empirical evaluation of our lifted tree reweighted (LTRW) algorithm. As a baseline we use a dampened version of the lifted belief propagation (LBP-Dampening) algorithm from [21]. Our lifted algorithm has all of the same advantages of the tree-reweighted approach over belief propagation, which we will illustrate in the results: (1) a convex objective that can be convergently solved to optimality, (2) upper bounds on the partition function, and (3) the ability to easily improve the approximation by tightening the relaxation. Our evaluation includes four variants of the LTRW algorithm corresponding to using different outer bounds: lifted local polytope (LTRW-L), lifted cycle polytope (LTRW-C), lifted local

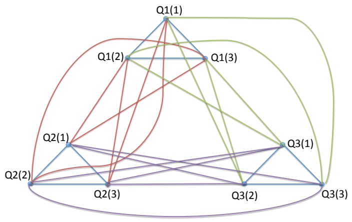

Figure 2: An example of the ground graphical model for the Clique-Cycle model (domain size = 3).

polytope with exchangeable polytope constraints (LTRW-LE), and lifted cycle polytope with exchangeable constraints (LTRW-CE). The conditional gradient optimization of the lifted TRW objective terminates when the duality gap is less than 10^{-4} or when a maximum number of 1000 iterations is reached. To solve the LP problem during conditional gradient, we use Gurobi[5].

We evaluate all the algorithms using several first-order probabilistic models. We assume that no evidence has been observed, which results in a large amount of symmetry. Even without evidence, performing marginal inference in first-order probabilistic models can be very useful for maximum likelihood learning [13]. Furthermore, the fact that our lifted tree-reweighted variational approximation provides an upper bound on the partition function enables us to maximize a lower bound on the likelihood [29], which we demonstrate in Sec. 7.5. To find the lifted orbit partition, we use the renaming group as in [3] which exploits the symmetry of the unobserved contants in the model.

Rather than optimize over the spanning tree polytope, which is computationally intensive, most TRW implementations use a single fixed choice of edge appearance probabilities, e.g. an (un)weighted distribution obtained using the matrix-tree theorem. In these experiments, we initialize the lifted edge appearance probabilities $\bar{\rho}$ to be the most uniform per-orbit edge-appearance probabilties by solving the optimization problem $\inf_{\bar{\rho}\in\mathbb{T}_{[\varphi E]}}(\bar{\rho}-\frac{|V|-1}{|E|})^2$ using conditional gradient. Each direction-finding step of this conditional gradient solves a lifted MST problem of the form $\sup_{\bar{\rho}'\in\mathbb{T}_{[\varphi E]}}\left\langle -2(\bar{\rho}-\frac{|V|-1}{|E|}),\bar{\rho}'\right\rangle$ using our lifted Kruskal's algorithm, where $\bar{\rho}$ is the current solution. After this initialization, we fix the lifted edge appearance probabilities and do not attempt to optimize them further.

7.1 Test models

Fig. 3 describes the four test models in MLN syntax. We focus on the repulsive case, since for attractive models, all TRW variants and lifted LBP give similar results. The parameter W denotes the weight that will be varying during the experiments. In all models except *Clique-Cycle*, W acts like the "local field" potential in an Ising model; a negative (or positive) value of W means the corresponding variable tends to be in the 0 (or 1) state. *Complete-Graph* is equivalent to an Ising model on the complete graph of size n (the domain size) with homogenous parameters. Exact marginals and the log-partition function can be computed in closed form using lifted counting elimination. The weight of the interaction clause is set to -0.1 (repulsive). *Friends-Smokers (negated)* is a variant of the Friends-Smokers model [21] where the weight of the final clause is set to -1.1 (repulsive). We use the method in [2] to compute the exact marginal for the *Cancer* predicate and the exact value of the log-partition function. *Lovers-Smokers* is the same MLN used in [3] with a full transitive clause and where we vary the prior of the *Loves* predicate. *Clique-Cycle* is a model with 3 cliques and 3 bipartite graphs in between. Its corresponding ground graphical model is shown in Fig. 2.

7.2 Accuracy of Marginals

Fig. 4 shows the marginals computed by all the algorithms as well as exact marginals on the Complete-Graph and Friends-Smokers models. We do not know how to efficiently perform exact inference in the remaining two models, and thus do not measure accuracy for them. The result on complete graphs illustrates the clear benefit of tightening the relaxation: LTRW-Local and LBP are inaccurate for moderate W, whereas cycle constraints and, especially, exchangeable constraints drastically improve accuracy. As discussed earlier, for the case of symmetric complete graphical models, the exchangeable constraints suffice to exactly characterize the marginal polytope. As a result, the approximate marginals computed by LTRW-LE and LTRW-CE are almost the same as the exact marginals; the very small difference is due to the entropy approximation. On the Friends-Smokers (negated) model, all LTRW variants give accurate marginals while lifted LBP even with very strong dampening (0.9 weight given to previous iterations' messages) fails to converge for $W<2$. We observed that LTRW-LE gives the best trade-off between accuracy and running time for this model. Note that we do not compare to ground versions of the lifted TRW algorithms because, by Theorem 3, the marginals and log-partition function are the same for both.

7.3 Quality of Log-Partition Upper bounds

Fig. 5 plots the values of the upper bounds obtained by the LTRW algorithms on the four test models. The results clearly show the benefits of adding each type of constraint to the LTRW, with the best upper bound obtained by tightening the lifted local polytope with both lifted cycle and exchangeable constraints. For the Complete-Graph and Friends-Smokers model, the log-partition approximation using exchangeable polytope constraints is very close to exact. In addition, we illustrate lifted LBP's approximation of the log-partition function on the Complete-Graph (note it is non-convex and not an upper bound).

[5] http://www.gurobi.com/

Complete Graph

W	$V(x)$
-0.1	$[x \neq y \wedge (V(x) \Leftrightarrow V(y))]$

Friends-Smokers (Negated)

W	$[x \neq y \wedge \neg Friends(x,y)]$
1.4	$\neg Smokes(x)$
2.3	$\neg Cancer(x)$
1.5	$Smokes(x) \Rightarrow Cancer(x)$
-1.1	$[x \neq y \wedge Smokes(x) \wedge Friends(x,y) \Rightarrow Smokes(y)]$

Lovers-Smokers

W	$[x \neq y \wedge Loves(x,y)]$
100	$Male(x) \Leftrightarrow !Female(x)$
2	$Male(x) \wedge Smokes(x)$
1	$Female(x) \wedge Smokes(x)$
0.5	$[x \neq y \wedge Male(x) \wedge Female(y) \wedge Loves(x,y)]$
1	$[x \neq y \wedge Loves(x,y) \wedge (Smokes(x) \Leftrightarrow Smokes(y))]$
-100	$[x \neq y \wedge y \neq z \wedge z \neq x \wedge Loves(x,y) \wedge Loves(y,z) \wedge Loves(x,z)]$

Clique-Cycle

W	$x \neq y \wedge (Q1(x) \Leftrightarrow \neg Q2(y))$
W	$x \neq y \wedge (Q2(x) \Leftrightarrow \neg Q3(y))$
W	$x \neq y \wedge (Q3(x) \Leftrightarrow \neg Q1(y))$
$-W$	$x \neq y \wedge (Q1(x) \Leftrightarrow Q1(y))$
$-W$	$x \neq y \wedge (Q2(x) \Leftrightarrow Q2(y))$
$-W$	$x \neq y \wedge (Q3(x) \Leftrightarrow Q3(y))$

Figure 3: Test models

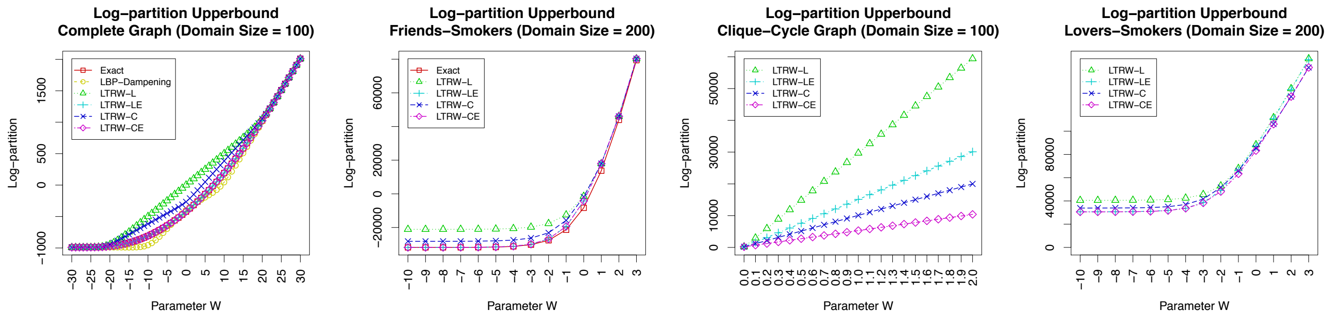

Figure 5: Approximations of the log-partition function on the four test models from Fig. 3 (best viewed in color).

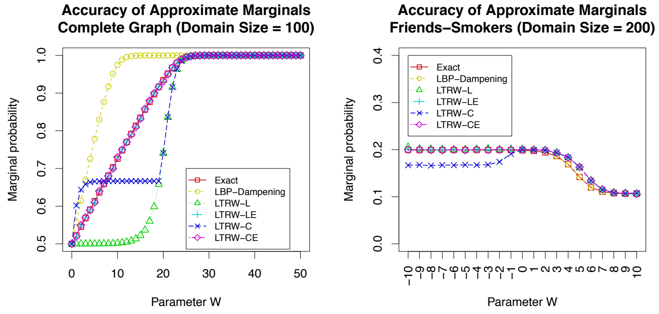

Figure 4: Left: marginal accuracy for complete graph model. Right: marginal accuracy for $Pr(Cancer(x))$ in Friends-Smokers (neg). Lifted TRW variants using different outer bounds: L=local, C=cycle, LE=local+exchangeable, CE=cycle+exchangeable (best viewed in color).

7.4 Running time

As shown in Table 1, lifted variants of TRW are order-of-magnitudes faster than the ground version. Interestingly, lifted TRW with local constraints is observed to be faster as the domain size increase; this is probably due to the fact that as the domain size increases, the distribution becomes more peak, so marginal inference becomes more similar to MAP inference. Lifted TRW with local and exchangeable constraints requires a smaller number of conditional gradient iterations, thus is faster; however note that its running time slowly increases since the exchangeable constraint set grows linearly with domain size.

LBP's lack of convergence makes it difficult to have a meaningful timing comparison with LBP. For example, LBP did not converge for about half of the values of W in the *Lovers-Smokers* model, even after using very strong dampening. We did observe that when LBP converges, it is much faster than LTRW. We hypothesize that this is due to the message passing nature of LBP, which is based on a fixed point update whereas our algorithm is based on Frank-Wolfe.

Domain size	10	20	30	100	200
TRW-L	138370	609502	1525140	-	-
LTRW-L	3255	3581	3438	1626	1416
LTRW-LE	681	703	721	1033	1307

Table 1: Ground vs lifted TRW runtime on Complete-Graph (milliseconds)

7.5 Application to Learning

We now describe an application of our algorithm to the task of learning relational Markov networks for inferring protein-protein interactions from noisy, high-throughput, experimental assays [12]. This is equivalent to learning the parameters of an exponential family random graph model [19] where edges in the random graph represent the protein-protein interactions. Despite fully observed data, maximum likelihood learning is challenging because of the intractability of computing the log-partition function and its gradient. In particular, this relational Markov network has over 330K random variables (one for each possible interaction of 813 variables) and tertiary potentials. However, Jaimovich et al. [13] observed that the partition function in

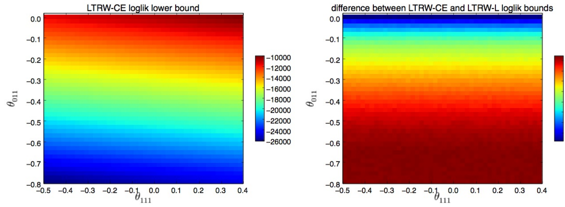

Figure 6: Log-likelihood lower-bound obtained using lifted TRW with the cycle and exchangeable constraints (CE) for the same protein-protein interaction data used in [13] (left) (c.f. Fig. 7 in [13]). Improvement in lower-bound after tightening the local constraints (L) with CE (right).

relational Markov networks is highly symmetric, and use lifted LBP to efficiently perform approximate learning in running time that is independent of the domain size. They use their lifted inference algorithm to visualize the (approximate) likelihood landscape for different values of the parameters, which among other uses characterizes the robustness of the model to parameter changes.

We use precisely the same procedure as [13], substituting lifted BP with our new lifted TRW algorithms. The model has three parameters: θ_1, used in the single-node potential to specify the prior probability of a protein-protein interaction; θ_{111}, part of the tertiary potentials which encourages cliques of three interacting proteins; and θ_{011}, also part of the tertiary potentials which encourages chain-like structures where proteins A, B interact, B, C interact, but A and C do not (see supplementary material for the full model specification as an MLN). We follow their two-step estimation procedure, first estimating θ_1 in the absence of the other parameters (the maximum likelihood, BP, and TRW estimates of this parameter coincide, and estimation can be performed in closed-form: $\theta_1^* = -5.293$). Next, for each setting of θ_{111} and θ_{011} we estimate the log-partition function using lifted TRW with the cycle+exchangeable vs. local constraints only. Since TRW is an upper bound on the log-partition function, these provide lower bounds on the likelihood.

Our results are shown in Fig. 6, and should be compared to Fig. 7 of [13]. The overall shape of the likelihood landscapes are similar. However, the lifted LBP estimates of the likelihood have several local optima, which cause gradient-based learning with lifted LBP to reach different solutions depending on the initial setting of the parameters. In contrast, since TRW is convex, any gradient-based procedure would reach the global optima, and thus learning is much easier. Interestingly, we see that our estimates of the likelihood have a significantly smaller range over these parameter settings than that estimated by lifted LBP. Moreover, the high-likelihood parameter settings extends to larger values of θ_{111}. For all algorithms there is a sudden decrease in the likelihood at $\theta_{011} > 0$ (not shown in the figure).

8 Discussion and Conclusion

Lifting partitions used by lifted and counting BP [21, 16] can be coarser than orbit partitions. In graph-theoretic terms, these partitions are called *equitable* partitions. If each equitable partition cell is thought of as a distinct node color, then among nodes with the same color, their neighbors must have the same color histogram. It is known that orbit partitions are always equitable, however the converse is not always true [9].

Since equitable partition can be computed more efficiently and potentially leads to more compact lifted problems, the following question naturally arises: can we use equitable partition in lifting the TRW problem? Unfortunately, a complete answer is non-trivial. We point out here a theoretical barrier due to the interplay between the spanning tree polytope and the equitable partition of a graph.

Let ε be the coarsest equitable partition of edges of $\mathcal{G}$. We give an example graph in the supplementary material (see example 9) where the symmetrized spanning tree polytope corresponding to the equitable partition ε, $\mathbb{T}_{[\epsilon]} = \mathbb{T}(\mathcal{G}) \cap \mathbb{R}^{|E|}_{[\varepsilon]}$ is an empty set. When $\mathbb{T}_{[\epsilon]}$ is empty, the consequence is that if we want ρ to be within $\mathbb{T}$ so that $B(., \rho)$ is guaranteed to be a convex upper bound of the log-partition function, we cannot restrict ρ to be consistent with the equitable partition. In lifted and counting BP, $\rho \equiv 1$ so it is clearly consistent with the equitable partition; however, one loses convexity and upper bound guarantee as a result. This suggests that there might be a trade-off between the compactness of the lifting partition and the quality of the entropy approximation, a topic deserving the attention of future work.

In summary, we presented a formalization of lifted marginal inference as a convex optimization problem and showed that it can be efficiently solved using a Frank-Wolfe algorithm. Compared to previous lifted variational inference algorithms, in particular lifted belief propagation, our approach comes with convergence guarantees, upper bounds on the partition function, and the ability to improve the approximation (e.g. by introducing additional constraints) at the cost of small additional running time.

A limitation of our lifting method is that as the amount of soft evidence (the number of distinct individual objects) approaches the domain size, the behavior of lifted inference approaches ground inference. The wide difference in running time between ground and lifted inference suggests that significant efficiency can be gained by solving an approximation of the orignal problem that is more symmetric [25, 15, 22, 6]. One of the most interesting open questions raised by our work is how to use the variational formulation to perform approxiate lifting. Since our lifted TRW algorithm provides an upper bound on the partition function, it is possible that one could use the upper bound to guide the choice of approximation when deciding how to re-introduce symmetry into an inference task.

Acknowledgements: Work by DS supported by DARPA PPAML program under AFRL contract no. FA8750-14-C-0005.

References

[1] F. Barahona and A. R. Mahjoub. On the cut polytope. *Mathematical Programming*, 36:157–173, 1986.

[2] Hung Hai Bui, Tuyen N. Huynh, and Rodrigo de Salvo Braz. Lifted inference with distinct soft evidence on every object. In *AAAI-2012*, 2012.

[3] Hung Hai Bui, Tuyen N. Huynh, and Sebastian Riedel. Automorphism groups of graphical models and lifted variational inference. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI-2013*. AUAI Press, 2013.

[4] Hung Hai Bui, Tuyen N. Huynh, and David Sontag. Lifted tree-reweighted variational inference. *arXiv*, 2014.

[5] R. de Salvo Braz, E. Amir, and D. Roth. Lifted first-order probabilistic inference. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI '05)*, pages 1319–1125, 2005.

[6] Rodrigo de Salvo Braz, Sriraam Natarajan, Hung Bui, Jude Shavlik, and Stuart Russell. Anytime lifted belief propagation. In *6th International Workshop on Statistical Relational Learning (SRL 2009)*, 2009.

[7] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. ISSN 1931-9193.

[8] A. Globerson and T. Jaakkola. Convergent Propagation Algorithms via Oriented Trees. In *Uncertainty in Artificial Intelligence*, 2007.

[9] Chris Godsil and Gordon Royle. *Algebraic Graph Theory*. Springer, 2001.

[10] Vibhav Gogate and Pedro Domingos. Probabilistic theorem proving. In *Proceedings of the Twenty-Seventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 256–265, 2011.

[11] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th ICML*, volume 28, pages 427–435. JMLR Workshop and Conference Proceedings, 2013.

[12] Ariel Jaimovich, Gal Elidan, Hanah Margalit, and Nir Friedman. Towards an integrated protein-protein interaction network: a relational markov network approach. *Journal of Computational Biology*, 13(2):145–164, 2006.

[13] Ariel Jaimovich, Ofer Meshi, and Nir Friedman. Template based inference in symmetric relational markov random fields. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19-22, 2007*, pages 191–199. AUAI Press, 2007.

[14] Jeremy Jancsary and Gerald Matz. Convergent decomposition solvers for tree-reweighted free energies. *Journal of Machine Learning Research - Proceedings Track*, 15:388–398, 2011.

[15] K. Kersting, Y. El Massaoudi, B. Ahmadi, and F. Hadiji. Informed lifting for message–passing. In D. Poole M. Fox, editor, *Twenty–Fourth AAAI Conference on Artificial Intelligence (AAAI–10)*, Atlanta, USA, July 11 – 15 2010. AAAI Press.

[16] Kristian Kersting, Babak Ahmadi, and Sriraam Natarajan. Counting belief propagation. In *Proceedings of the 25th Annual Conference on Uncertainty in AI (UAI '09)*, 2009.

[17] B. Milch, L. S. Zettlemoyer, K. Kersting, M. Haimes, and L. P. Kaelbling. Lifted Probabilistic Inference with Counting Formulas. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI '08)*, pages 1062–1068, 2008.

[18] Mathias Niepert. Markov chains on orbits of permutation groups. In *UAI-2012*, 2012.

[19] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p^*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[20] H. D. Sherali and W. P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3 (3):411–430, 1990. doi: 10.1137/0403036. URL `http://link.aip.org/link/?SJD/3/411/1`.

[21] Parag Singla and Pedro Domingos. Lifted first-order belief propagation. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI '08)*, pages 1094–1099, 2008.

[22] Parag Singla, Aniruddh Nath, and Pedro Domingos. Approximate lifted belief propagation. In *Workshop on Statistical Relational Artificial Intelligence (StaR-AI 2010)*, 2010.

[23] D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In *Advances in Neural Information Processing Systems 21*. MIT Press, 2008.

[24] David Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *Proceedings of the 24th Annual Conference on Uncertainty in AI (UAI '08)*, 2008.

[25] Guy Van den Broeck and Adnan Darwiche. On the complexity and approximation of binary evidence in lifted inference. In *Advances in Neural Information Processing Systems*, pages 2868–2876, 2013.

[26] M. Wainwright, T. Jaakkola, and A. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51:2313–2335, 2005.

[27] Martin Wainwright and Michael Jordan. *Graphical Models, Exponential Families, and Variational Inference.* Now Publishers, 2008.

[28] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory*, 49:1120–1146, 2003.

[29] C. Yanover, O. Schueler-Furman, and Y. Weiss. Minimizing and learning energy functions for side-chain prediction. *Journal of Computational Biology*, 15(7):899–911, 2008.