
Lifted Weight Learning of Markov Logic Networks Revisited

Ondřej Kuželka

Czech Technical University in Prague
Dept of Computer Science, KU Leuven

Vyacheslav Kungurtsev

Czech Technical University in Prague

Abstract

We study lifted weight learning of Markov logic networks. We show that there is an algorithm for maximum-likelihood learning of 2-variable Markov logic networks which runs in time polynomial in the domain size. Our results are based on existing lifted-inference algorithms and recent algorithmic results on computing maximum entropy distributions.

1 INTRODUCTION

Statistical Relational Learning [7] (SRL) is concerned with learning probabilistic models from relational data. Markov Logic Networks [9] (MLNs) are among the most prominent SRL systems. An MLN is given by a set of weighted first-order logic formulas and a domain Δ . Generative weight learning of MLNs is typically performed using maximum-likelihood estimation. Unfortunately, maximizing likelihood of MLNs is generally intractable. Therefore, in practice, one often resorts to heuristic approximations. Another option besides using approximations is to restrict the class of MLNs to those for which inference can be performed efficiently. This has been studied in the subarea of SRL called *lifted inference* [5]. In particular, it has been shown in [13, 12] that probabilistic inference in MLNs with formulas containing at most 2 logical variables can be performed in time polynomial in the size of the given domain Δ . This has been exploited in [14] for maximum-likelihood learning of MLNs, suggesting tractable learning of 2-variable MLNs could be possible. However, although it showed that gradients of log-likelihood can be computed efficiently, it did not provide a bound on the total runtime of the learning algorithm, specifically, because this bound was miss-

ing a guarantee on the number of iterations of the optimization algorithm.

In this paper, we complete the work of [14] by answering whether maximum-likelihood learning of MLNs can be done in time polynomial in the size of the domain for 2-variable MLNs. We give a positive answer to this question (Theorem 11), under consideration of the dependence of the runtime bounds on how extreme the statistics of the training data are. To arrive at this positive result, we need to combine results from three streams of research: (i) lifted inference [13, 2], (ii) links between maximum-likelihood learning of MLNs and relational marginal problems [8], and (iii) algorithmic results on maximum-entropy distributions [11]. We should note here that our results are mostly of theoretical interest. Making the algorithms described in this paper practical would be potential future research.

The rest of the paper is structured as follows. After covering the necessary background material in Section 2, we introduce the concept of interiority in relational marginal polytopes in Section 3. We then state our main technical results in Section 4. Then, in Sections 5, 6, we work towards the proof of the main results which we finish in Sections 7 and 8. The paper is concluded in Section 9.

2 BACKGROUND

2.1 First-Order Logic

We consider a function-free first-order logic language \mathcal{L} , built from a set of constants $Const$, variables Var and predicates $Rel = \bigcup_i Rel_i$, where Rel_i contains the predicates of arity i . We assume an untyped language (all our results can be straightforwardly generalized to the typed case). For $a_1, \dots, a_k \in Const \cup Var$ and $R \in Rel_k$, we call $R(a_1, \dots, a_k)$ an atom. If $a_1, \dots, a_k \in Const$, this atom is called ground. A literal is an atom or its negation. We use $Vars(\alpha)$ to denote the variables that appear in a formula α . The formula α_0 is called a grounding of α if α_0 can be obtained by replacing each variable in α with a constant from $Const$. A formula is

called closed if all variables are bound by a quantifier. A variable in a formula is called free if it is not bound by a quantifier. A formula with no free variables is called a sentence. A formula is called quantifier-free if all variables in it are free. A possible world ω is defined as a set of ground atoms. A substitution is a mapping from variables to terms. An injective substitution is a substitution which does not map any two variables to the same variable or constant.

2.2 Markov Logic Networks

A Markov logic network [9] (MLN) is a set of weighted first-order logic formulas (α, w) , where $w \in \mathbb{R}$ and α is a function-free and quantifier-free first-order formula. The semantics are defined w.r.t. the groundings of the first-order formulas, relative to some finite set of constants Δ , called the domain. An MLN is classically seen as a template that defines a Markov random field (in Section 2.4, we describe another way of interpreting MLNs—as solutions to max-entropy marginal problems). Specifically, an MLN Φ induces the following probability distribution on the set of possible worlds $\omega \in \Omega$: $p_\Phi(\omega) = \frac{1}{Z} \exp\left(\sum_{(\alpha, w) \in \Phi} w \cdot N(\alpha, \omega)\right)$, where $N(\alpha, \omega)$ is the number of injective¹ groundings of α satisfied in ω , and Z is a normalization constant to ensure that p_Φ is a probability distribution.

2.3 Ellipsoid Algorithm

In this section we briefly describe the main properties of the ellipsoid algorithm for convex optimization [4]; the exposition is based on [11]. Consider an arbitrary convex optimization problem,

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & g(\lambda) \\ \text{s.t.} \quad & h_i(\lambda) = 0, \forall i \in \{1, \dots, k\} \end{aligned}$$

where g is concave and h_i are all affine. Assume that g and h_i are differentiable everywhere, and furthermore, there exists a *strong first order oracle* for g which, given λ , outputs $g(\lambda)$ and $\nabla g(\lambda)$ and that we can project $\nabla g(\lambda)$ onto the affine space defined by $K = \{\lambda : h_i(\lambda) = 0, \forall i \in \{1, \dots, k\}\}$.

The ellipsoid algorithm will be used in the proofs in this paper as it satisfies the following property,

Theorem 1. [11, Theorem 2.13] *Given any $\beta, R > 0$, there exists an algorithm, namely the ellipsoid algorithm which, given a strong first-order oracle for g , returns a $\hat{\lambda}$ such that,*

$$g(\hat{\lambda}) \geq \max_{\lambda \in K, \|\lambda\|_\infty \leq R} g(\lambda)$$

¹Normally, MLNs are not defined with injective groundings. However, working with injective groundings turns out to be more convenient and equally expressive [8, 6].

$$+ \beta \left(\min_{\lambda \in K, \|\lambda\|_\infty \leq R} g(\lambda) - \max_{\lambda \in K, \|\lambda\|_\infty \leq R} g(\lambda) \right)$$

and the number of calls to the strong first-order oracle is bounded by a polynomial in m , $\log R$ and $\log(1/\beta)$.

2.4 Relational Marginal Problems

In this section we describe the relationship between MLN weight learning using maximum likelihood estimation and so-called *relational marginal problems* which were studied in [8].

We start by defining *formula statistics* which are closely related to *random-substitution semantics* [1, 10]. In our case, the formula statistics are just rescaled numbers of true groundings of a formula (defined in Section 2.2), where the scaling depends on the number of variables in the formula.


Definition 1 (Formula statistics). *Let α be a quantifier-free first-order logic formula with k variables $\{x_1, \dots, x_k\}$. We define its formula statistic w.r.t. a possible world ω as:*

$$Q_\omega(\alpha) = \binom{|\Delta|}{k}^{-1} \cdot (k!)^{-1} \cdot N(\alpha, \omega).$$

Remark 2. *When α does not contain any variables, e.g. when $\alpha = \text{smokes}(\text{Alice})$, then $Q_\omega(\alpha) \in \{0, 1\}$.*

Intuitively, for a given formula α and a possible world ω , the formula statistic $Q_\omega(\alpha)$ is the probability that the ground formula $\alpha\vartheta$ is true where ϑ is a grounding injective substitution of α 's free variables picked from all such substitutions uniformly at random.²

Example 3. *Let $\omega = \{\text{fr}(\text{Alice}, \text{Bob}), \text{fr}(\text{Bob}, \text{Alice}), \text{fr}(\text{Bob}, \text{Eve}), \text{fr}(\text{Eve}, \text{Bob}), \text{sm}(\text{Alice})\}$ and $\Delta = \{\text{Alice}, \text{Bob}, \text{Eve}\}$, i.e. the only smoker is Alice and the*

friendship structure is:  *Let $\alpha = \text{fr}(x, y) \Rightarrow \text{sm}(y)$. We then get $Q_\omega(\alpha) = \frac{1}{2}$ (of the 6 possible injective substitutions ϑ of α 's variables, three make $\alpha\vartheta$ true in ω).*

Remark 4. *Let us have a set Ω of possible worlds over a domain Δ . MLNs over Ω , given by a set of weighted formulas $\Phi = \{(\alpha_1, w_1), \dots, (\alpha_l, w_l)\}$, can be re-defined using formula statistics as:*

$$p_\Phi(\omega) = \frac{1}{Z} \exp \left(\sum_{(\alpha_i, w_i) \in \Phi} w_i \cdot Q_\omega(\alpha_i) \right).$$

For possible worlds over a domain of fixed size, the only difference is the scaling factor in the definition of formula statistics, which is fixed for each formula and

²This is how formula statistics relate to random substitution semantics [1, 10].

fixed domain size, hence, as a result the only difference is that the weights need to be scaled as well. In what follows when we refer to MLNs we will mean this representation unless stated otherwise.

Next we use formula statistics to define a maximum entropy distribution over a given set of possible worlds Ω . Assuming that we know the values $\theta_1, \dots, \theta_l$ that the formula statistics of the given formulas $\alpha_1, \dots, \alpha_l$ should have in expectation (which we might have, for instance, estimated from given training data), we can define the following convex optimization problem encoding the maximum entropy problem.

Relational Marginal Problem (Formulation):

$$\min_{\{P_\omega : \omega \in \Omega\}} \sum_{\omega \in \Omega} P_\omega \log P_\omega \quad s.t. \quad (1)$$

$$\forall i = 1, \dots, l : \sum_{\omega \in \Omega} P_\omega \cdot Q_\omega(\alpha_i) = \theta_i \quad (2)$$

$$\forall \omega \in \Omega : P_\omega \geq 0, \sum_{\omega \in \Omega} P_\omega = 1 \quad (3)$$

Here, P_ω 's are the decision variables of the problem, each representing probability of one possible world $\omega \in \Omega$. The first line (1) is the maximum entropy criterion (represented here as minimization of negative entropy), (2) are constraints given by the statistics and (3) are normalization constraints for the probability distribution.

Assuming there exists a feasible solution satisfying $\forall \omega : P_\omega > 0$, the optimal solution of the above maximum entropy problem is an MLN

$$P_\omega = p_\Phi(\omega) = \frac{1}{Z} \exp \left(\sum_{(\alpha_i, \lambda_i) \in \Phi} \lambda_i \cdot Q_\omega(\alpha_i) \right) \quad (4)$$

where the parameters $\lambda = (\lambda_1, \dots, \lambda_l)$ are obtained by maximizing the dual criterion

$$L(\lambda) = \sum_{\alpha_i} \lambda_i \theta_i - \log \sum_{\omega \in \Omega} e^{\sum_{\alpha_i} \lambda_i Q_\omega(\alpha_i)} \quad (5)$$

This dual criterion also happens to be equivalent to the log-likelihood of the MLN (4) w.r.t. a (possibly fictitious) training example $\hat{\omega}$ that has to be over the same domain Δ and that satisfies $Q_{\hat{\omega}}(\alpha_i) = \theta_i$ for all the formula statistics.

Remark 5. *Due to the above duality, if we can show that we can solve relational marginal problems efficiently, it will follow as a corollary that we can solve maximum likelihood estimation in MLNs efficiently and vice versa.*

Remark 6. *Above, we have used the assumption that there exists a feasible solution where probability of every possible world is positive. This does not hurt generality of our discussion because we can always remove the possible worlds ω that, by the virtue of the given constraints, must have zero probability in any feasible solution from the set Ω . In most cases, Ω is not given explicitly but by means of a first-order logic theory (that describes which possible worlds are “possible”), so it is enough to add suitable first-order sentences to this theory.*

2.5 Inference Using Weighted Model Counting

To maximize the dual criterion (5) we will need to be able to compute its gradient. For the partial derivatives of (5), we have

$$\begin{aligned} \frac{\partial L}{\partial \lambda_i} &= \theta_i - \frac{\sum_{\omega \in \Omega} Q_\omega(\alpha_i) \cdot e^{\sum_{\alpha_j} \lambda_j Q_\omega(\alpha_j)}}{\sum_{\omega \in \Omega} e^{\sum_{\alpha_j} \lambda_j Q_\omega(\alpha_j)}} \\ &= \theta_i - \mathbb{E}[Q_\omega(\alpha_i)] \end{aligned} \quad (6)$$

It follows that, in order to compute the gradient, we will also need to be able to compute the partition function $Z = \sum_{\omega \in \Omega} e^{\sum_{\alpha_j} \lambda_j Q_\omega(\alpha_j)}$. Computation of the partition function Z can be converted to a *first-order weighted model counting problem (WFOMC)*.

Definition 2 (WFOMC [13]). *Let $w(P)$ and $\bar{w}(P)$ be functions from predicates to real numbers (we call w and \bar{w} weight functions) and let Φ be a first-order theory. Then $\text{WFOMC}(\Phi, w, \bar{w}) =$*

$$= \sum_{\omega \in \Omega : \omega \models \Phi} \prod_{a \in \mathcal{P}(\omega)} w(\text{Pred}(a)) \prod_{a \in \mathcal{N}(\omega)} \bar{w}(\text{Pred}(a))$$

where $\mathcal{P}(\omega)$ and $\mathcal{N}(\omega)$ denote the positive literals that are true and false in ω , respectively, and $\text{Pred}(a)$ denotes the predicate of a (e.g. $\text{Pred}(\text{friends}(\text{Alice}, \text{Bob})) = \text{friends}$).

To compute the partition function Z using weighted model counting, we may proceed as in [13]. Let a set of weighted formulas Φ be given. Here, for simplicity of exposition, we will assume that the formulas in Φ do not contain constants (we refer to [13] for the general case). For every weighted formula $(\alpha_i, \lambda_i) \in \Phi$, where the free variables in α_i are exactly x_1, \dots, x_k , we create a new formula

$$\forall x_1, \dots, x_k : \xi_i(x_1, \dots, x_k) \Leftrightarrow (\alpha_i(x_1, \dots, x_k) \wedge x_1 \neq x_2 \wedge x_1 \neq x_3 \wedge \dots \wedge x_{k-1} \neq x_k)$$

where ξ is a new fresh predicate. Then we set

$$w(\xi_i) = \exp \left(\left(\frac{|\Delta|}{|\text{Vars}(\alpha_i)|} \right)^{-1} \cdot (|\text{Vars}(\alpha_i)|!)^{-1} \cdot \lambda_i \right)$$

and $\bar{w}(\xi_i) = 1$ and for all other predicates we set both w and \bar{w} equal to 1. It is easy to check that then $WFOMC(\Phi, w, \bar{w}) = Z$, which is what we needed to compute. To compute the numerator of (6), we need to compute $WFOMC(\Phi \cup \{\alpha_i \vartheta\}, w, \bar{w})$ where ϑ is an injective grounding substitution of α_i .

Importantly, there are classes of first-order logic theories for which weighted model counting is polynomial-time. In particular, as shown in [12], when the theory consists only of first-order logic sentences, each of which contains at most two logic variables, the weighted model count can be computed in time polynomial³ in the number of elements in the domain Δ over which the set of possible worlds Ω is defined. This is not the case in general when the number of variables in the formulas is greater than two unless $P = \#P_1$ [2].

Remark 7. *It has already been shown in [14] that gradients of log-likelihood of an MLN can be computed efficiently whenever WFOMC can be computed efficiently (in fact, the translation described in this section for computing Z is essentially the same as the one described in [14]).*

3 MARGINAL POLYTOPES

Not all possible values of formula statistics correspond to actual probability distributions.

Example 8. *Let $\alpha = e(x_1, x_2)$, $\beta = e(x_1, x_2) \wedge e(x_2, x_3) \wedge e(x_3, x_1)$ and let $\Delta = \{c_1, \dots, c_{100}\}$ be the set of domain elements and Ω be the respective set of possible worlds over the first-order language given by the predicate $e/2$ and the constants from Δ . We can think of possible worlds $\omega \in \Omega$ as directed graphs (the predicate $e/2$ representing edges in the graph and the constants in Δ representing vertices). Then $Q_\omega(\alpha)$ corresponds to “density” of edges and $Q_\omega(\beta)$ to “density” of directed triangles. It is then easy to see why there is, for instance, no distribution with $\mathbb{E}[Q_\omega(\alpha)] = 0$ and $\mathbb{E}[Q_\omega(\beta)] = 0.5$ (since graphs with no edges obviously cannot have positive density of triangles).*

The points corresponding to values of statistics that correspond to some actual probability distributions form what is called a *relational marginal polytope* [8].

Definition 3 (Relational marginal polytope). *Let Ω be a set of possible worlds and $\Phi = (\alpha_1, \dots, \alpha_l)$ be a list of formulas. We define the relational marginal polytope $RMP(\Phi, \Omega)$ w.r.t. Φ as*

$$RMP(\Phi, \Omega) = \{(x_1, \dots, x_l) \in R^l : \exists \text{ prob. distr. on } \Omega \text{ s.t. } \mathbb{E}[Q_\omega(\alpha_i)] = x_i\}.$$

³Here, we should note that the runtime of these WFOMC algorithms depends on the parameters of the theory Φ exponentially. However, in many cases, these parameters are small compared to size of the domain.

$$\Omega \text{ s.t. } \mathbb{E}[Q_\omega(\alpha_1)] = x_1 \wedge \dots \wedge \mathbb{E}[Q_\omega(\alpha_l)] = x_l\}.$$

Remark 9. *It is not difficult to see that the relational marginal polytope w.r.t. a given list of formulas $(\alpha_1, \dots, \alpha_l)$ can be equivalently defined as the convex hull of the set $\{(Q_\omega(\alpha_1), \dots, Q_\omega(\alpha_l)) : \omega \in \Omega\}$.*

Next we define what it means for a point to be in the η -interior of a polytope.

Definition 4 (Interiority). *Let $\eta > 0$, \mathbf{P} be a polytope and $A\bar{\mathbf{x}} = \mathbf{c}$ be the maximal linearly independent system of linear equations that hold for the vertices of \mathbf{P} . A point θ is said to be in the η -interior of \mathbf{P} if $\{\theta' | A\bar{\theta}' = \mathbf{c}, \|\theta' - \theta\| \leq \eta\} \subseteq \mathbf{P}$.*

The reason why we need to consider the system of linear equations $A\bar{\mathbf{x}} = \mathbf{c}$ in the definition of interiority is because it may happen that the polytope lives in a lower dimensional subset of the given space. We note that interiority, as we defined it, is also often called *relative interiority* in the literature.

Remark 10. *When we were constructing the dual relational marginal problem, we had to assume that there is a positive solution that satisfies the constraints of the primal problem. It is not difficult to see that if the vector of formula statistics’ estimates θ is in the η -interior of the respective relational marginal polytope for some $\eta > 0$ then such a solution always exists. To see this, first, notice that an interior point θ can be represented as a convex combination $\theta = \sum_{\mathbf{x} \in \{(Q_\omega(\alpha_1), \dots, Q_\omega(\alpha_l)) : \omega \in \Omega\}} a_{\mathbf{x}} \cdot \mathbf{x}$ where $a_{\mathbf{x}} > 0$ for all $\mathbf{x} \in \{(Q_\omega(\alpha_1), \dots, Q_\omega(\alpha_l)) : \omega \in \Omega\}$. To find a positive distribution over Ω that satisfies the constraints, we just need to assign positive probabilities P_ω so that $a_{\mathbf{x}} = \sum_{\omega \in \Omega : (Q_\omega(\alpha_1), \dots, Q_\omega(\alpha_l)) = \mathbf{x}} P_\omega$, which we can always do.*

4 MAIN RESULTS

In this section we describe our main technical result which is showing that maximum-likelihood weight learning of 2-variable MLNs can be done in time polynomial in the size of the domain (i.e. the problem is domain-liftable [13]). As already mentioned in the previous sections, it has been shown that computing log-likelihood and its derivatives is domain liftable [13, 14] but it has not been shown what is the computational complexity of the complete weight learning problem.

It turns out that it is natural to study the complexity of the weight learning problem in the relational marginal setting because one of the parameters that influences runtime is interiority of the vectors which represent marginal constraints. In particular we have the following result which provides a polynomial-time

bound for maximum likelihood weight learning of 2-variable MLNs.

Theorem 11. *Let $\Phi = \{\alpha_1, \dots, \alpha_l\}$ be a set of quantifier-free first-order logic formulas, each with at most 2 variables. Let Φ_0 be a set of universally quantified first-order logic sentences, each also with at most 2 variables. Let Ω_{Φ_0} be the set of models of Φ_0 over a given domain Δ . Let $\hat{\omega} \in \Omega$ be a training example. Then there is an algorithm which finds weights of the MLN \mathcal{M} given by formulas Φ such that the log-likelihood of \mathcal{M} given the training example $\hat{\omega}$ is within ε of the optimum. The algorithm runs in time polynomial in $|\Delta|$, $1/\varepsilon$ and $1/\eta$ where η is the interiority of the vector $Q_{\hat{\omega}}(\Phi)$ in the relational marginal polytope $RMP(\Phi, \Omega_{\Phi_0})$.*

At first, one might perhaps wonder why the above result about maximum-likelihood estimation should depend on interiority of $Q_{\hat{\omega}}(\Phi)$. Consider the following example: $\hat{\omega}$ represents a complete directed graph (e.g. using binary relations $e/2$) and $\Phi = \{e(x, y)\}$. Then $Q_{\hat{\omega}}(\Phi) = (1)$ which is clearly on the boundary of the respective polytope (in this case the polytope is just a line segment). If we try to optimize likelihood of the MLN given by Φ , the weight of the formula $e(x, y)$ will tend to infinity which also means that the optimization algorithm will not be able to converge. Thus, some dependence on interiority is necessary.

While the case from the previous paragraph might be simple to spot, there are other more tricky cases where, at first, we might not be able to realize that the weights will have to be very large. For instance, consider MLNs given by two formulas, one for edge density and one for triangle density (as in Example 8). If the training example $\hat{\omega}$ turned out to represent a graph close to an extremal graph (see e.g. [3]), e.g. one having close to maximum possible density of triangles for the given density of edges, then the learned weights would again turn out to be very large, but this time because of a more subtle reason. Again, this is what η -interiority captures.

Finally, using Theorem 11, the duality of relational marginal problems and maximum-likelihood estimation in MLNs and a lemma from [11], we can obtain the next result about complexity of the relational marginal problems.

Theorem 12. *Let Φ , Φ_0 , Δ and Ω_{Φ_0} be as in Theorem 11 (in particular, all formulas in Φ and Φ_0 are still required to have at most 2 variables). Let $\eta > 0$ be a real number and $\theta = (\theta_1, \dots, \theta_l)$ be a point in the η -interior of the relational marginal polytope $RMP(\Phi, \Omega_{\Phi_0})$. Then there exists an algorithm which finds a distribution over Ω_{Φ_0} , represented as an MLN, whose entropy is within $\varepsilon > 0$ of the maximum and*

which satisfies the marginal constraints $\mathbb{E}(Q_{\omega}(\Phi)) = \theta$ within $\sqrt{\varepsilon}$. The runtime of this algorithm is polynomial in $|\Delta|$, $1/\eta$, $1/\varepsilon$ and the number of bits needed to represent θ .

Remark 13. *We have omitted using the term “domain-liftable” [13] in the description of the above two results. Here is why. Suppose that we fix a vector θ and increase the domain size $|\Delta|$. It can happen that θ becomes much closer to the boundary of the polytope which means that the runtime may increase more than just polynomially with increasing $|\Delta|$ because interiority of the vector θ is one of the parameters governing the runtime. In fact, θ may end up being completely outside the polytope, rendering the problem unsolvable. One possible solution is to use interiority w.r.t. the polytope that we obtain as a limit for $|\Delta| \rightarrow \infty$. It follows from results in [8] that polytopes over larger domains (but given by the same formulas Φ) are subsets of polytopes over smaller domains (one can also obtain bounds on how much smaller the limit polytope will be compared to some polytope over a finite domain using Proposition 8 in [8]). It follows that our results imply domain-liftability of the relational marginal problems for vectors θ that are in the interior of the respective limit polytopes (for $|\Delta| \rightarrow \infty$).*

We prove Theorem 11 and Theorem 12 in the next sections.

Outline of the Proof: First, we show how to construct relational marginal polytopes (which turn out to be needed by the algorithm) in Section 5. Then, in Section 6, following the approach from [11] we bound the weights of the MLN which is a solution of the relational marginal problem. We finish the rest of the proofs in Sections 7 and 8.

5 POLYTOPES FOR 2-VARIABLE FORMULAS

For our main result, a polynomial-time algorithm for solving relational marginal problems, we will need to be able to construct relational marginal polytopes in time polynomial in the size of the domain Δ . First, we may notice that the number of possible vectors of formulas’ statistics given by a fixed set of formulas can be bounded by a polynomial in Δ .

Remark 14. *Let $\Phi = (\alpha_1, \dots, \alpha_l)$ and let Ω be a set of possible worlds over a domain Δ . Let us define $\mathcal{K}(\Phi, \Omega) = \{(Q_{\omega}(\alpha_1), \dots, Q_{\omega}(\alpha_l)) \mid \omega \in \Omega\}$. Then $|\mathcal{K}(\Phi, \Omega)| \leq \prod_{\alpha_i \in \Phi} (|\Delta| + 1)^{|\text{Vars}(\alpha_i)|}$, which is polynomial in $|\Delta|$.*

Since the relational marginal polytope $RMP(\Phi, \Omega)$ is equal to the convex hull of $\mathcal{K}(\Phi, \Omega)$, the above remark

also provides a polynomial bound for the number of its vertices.

The next proposition is a consequence of an algorithm that we describe in the appendix.

Proposition 15. *Let Φ be a set of quantifier-free first-order logic formulas, each with at most 2 variables. Let Φ_0 be a set of universally quantified first-order logic sentences, each also with at most 2 variables. Finally, let Ω_{Φ_0} be the set of models of Φ_0 over a given domain Δ . Then the set of vertices of $RMP(\Phi, \Omega_{\Phi_0})$ can be constructed in time polynomial in $|\Delta|$.*

6 BOUNDING BOX

The main result described in this section is the following theorem which allows us to bound the magnitude of weights in MLNs that we obtain as solutions of relational marginal problems. This theorem is a relational counterpart of Theorem 2.7 from [11]. The proof follows the steps of the respective proof from [11] and most of the heavy-lifting has already been done there (however, we do need to generalize their results to our setting).

Theorem 16. *Let Φ be a set of quantifier-free first-order logic formulas, let Ω be a set of possible worlds and $A^\dagger \mathbf{x} = \mathbf{c}$ be a maximal system of linearly independent equations satisfied by the vertices of the relational marginal polytope $\mathbf{P}_R = RMP(\Phi, \Omega)$. Let θ be a point in the η -interior of \mathbf{P}_R . Then there is an optimal solution λ^* of the dual problem (5) such that $A^\dagger \lambda^* = 0$ and any such solution satisfies $\|\lambda^*\| \leq \log |\Omega| / \eta$.*

To prove this theorem we start with some lemmas. In what follows, when $\Phi = (\alpha_1, \dots, \alpha_l)$ is a list of formulas, we will use the notation $Q_\omega(\Phi) \triangleq (Q_\omega(\alpha_1), \dots, Q_\omega(\alpha_l))$.

Lemma 1.⁴ *Let $\Phi = (\alpha_1, \dots, \alpha_l)$, $\theta = (\theta_1, \dots, \theta_l)$ be a point in the η -interior of the relational marginal polytope $\mathbf{P}_R = RMP(\Phi, \Omega)$ and let $\lambda^* = (\lambda_1^*, \dots, \lambda_l^*)$ be the optimal solution to the dual problem (5). Then for any $\mathbf{x} \in \mathbf{P}_R$: $\langle \lambda^*, \mathbf{x} - \theta \rangle \leq \log |\Omega|$.*

Proof. The entropy of any distribution which is a solution of the relational marginal problem is bounded by $\log |\Omega|$, which is the entropy of the uniform distribution over Ω . It follows from strong duality that $-L(\lambda^*) \leq \log |\Omega|$ where $L(\lambda^*)$ is defined in (5). Hence

$$-L(\lambda^*) = -\langle \lambda^*, \theta \rangle + \log \sum_{\omega \in \Omega} e^{\langle \lambda^*, Q_\omega(\Phi) \rangle} \leq \log |\Omega|.$$

In particular, for every $\omega \in \Omega$:

$$-\langle \lambda^*, \theta \rangle + \langle \lambda^*, Q_\omega(\Phi) \rangle \leq \log |\Omega|. \quad (7)$$

⁴This is a relational counterpart of Lemma 5.1 from [11].

Since $x \in \mathbf{P}_R$, we can write it as a convex combination $x = \sum_{\omega \in \Omega} a_\omega \cdot Q_\omega(\Phi)$. Using (7) we obtain

$$\sum_{\omega \in \Omega} (-a_\omega \langle \lambda^*, \theta \rangle + a_\omega \langle \lambda^*, Q_\omega(\Phi) \rangle) \leq \sum_{\omega \in \Omega} a_\omega \log |\Omega|.$$

Since $\sum_{\omega \in \Omega} a_\omega = 1$ (recall that we represented \mathbf{x} as a convex combination), we obtain: $\langle \lambda^*, \mathbf{x} - \theta \rangle \leq \log |\Omega|$. \square

Lemma 2.⁵ *Let $A^\dagger \mathbf{x} = \mathbf{c}$ be a maximal linearly-independent system of linear equations which are satisfied by all vertices of the relational marginal polytope $\mathbf{P}_R = RMP(\Phi, \Omega)$. Then, for any $\mathbf{d} \in \mathbb{R}^m$ where m is the column dimension of A^\dagger , $L(\lambda) = L(\lambda + (A^\dagger)^T \mathbf{d})$ where L is as in (5).*

Proof. First, for any $\omega \in \Omega$: $A^\dagger Q_\omega(\Phi) = \mathbf{c}$. Second we can write $\theta = \sum_{\omega \in \Omega} a_\omega Q_\omega(\Phi)$, where $\sum_{\omega \in \Omega} a_\omega = 1$.

Next, we have

$$\begin{aligned} \langle \lambda + (A^\dagger)^T \mathbf{d}, \theta \rangle &= \langle \lambda, \theta \rangle + \langle (A^\dagger)^T \mathbf{d}, \theta \rangle \\ &= \langle \lambda, \theta \rangle + \sum_{\omega \in \Omega} a_\omega \langle (A^\dagger)^T \mathbf{d}, Q_\omega(\Phi) \rangle \\ &= \langle \lambda, \theta \rangle + \sum_{\omega \in \Omega} a_\omega \langle \mathbf{d}, A^\dagger Q_\omega(\Phi) \rangle = \langle \lambda, \theta \rangle + \langle \mathbf{d}, \mathbf{c} \rangle. \end{aligned}$$

For the dual problem (5), we have

$$\begin{aligned} L(\lambda + (A^\dagger)^T \mathbf{d}) &= \langle \lambda + (A^\dagger)^T \mathbf{d}, \theta \rangle \\ &\quad - \log \sum_{\omega \in \Omega} e^{\langle \lambda + (A^\dagger)^T \mathbf{d}, Q_\omega(\Phi) \rangle} = \langle \lambda, \theta \rangle + \langle \mathbf{d}, \mathbf{c} \rangle \\ &\quad - \log \sum_{\omega \in \Omega} e^{\langle \mathbf{d}, \mathbf{c} \rangle + \langle \lambda, Q_\omega(\Phi) \rangle} \\ &= \langle \lambda, \theta \rangle - \log \sum_{\omega \in \Omega} e^{\langle \lambda, Q_\omega(\Phi) \rangle} = L(\lambda). \end{aligned}$$

\square

Due to the above lemma and since A^\dagger represents the maximal set of linearly independent equalities satisfied by points of \mathbf{P}_R , we can restrict ourselves to λ 's that satisfy $A^\dagger \lambda = 0$ in the search for the optimal solution of the dual problem (5).

The next lemma, which we will also need for the proof of Theorem 16, does not need to be adapted and can be used for our purposes as is; we refer to [11] for proof.

Lemma 3 (Lemma 5.2 in [11]). *Let $A^\dagger \mathbf{x} = \mathbf{c}$ be a system of linear equations, $\theta \in \mathbb{R}^m$ and $\eta \geq 0$. Let us define three sets \mathcal{B} , \mathcal{Q} and $\tilde{\mathcal{Q}}$:*

$$\mathcal{B}(\theta) = \{\mathbf{x} \in \mathbb{R}^m \mid A^\dagger \mathbf{x} = \mathbf{c}, \|\mathbf{x} - \theta\| \leq \eta\},$$

⁵This is a relational counterpart of Lemma 2.5 in [11].

$$\begin{aligned} \mathcal{Q}(\theta) &= \{\mathbf{y} \in \mathbb{R}^m \mid A^\top \mathbf{y} = \mathbf{c}, \|\mathbf{y} - \theta\| \leq 1/\eta\}, \\ \tilde{\mathcal{Q}}(\theta) &= \{\mathbf{z} \in \mathbb{R}^m \mid A^\top \mathbf{z} = \mathbf{c}, \forall x \in \mathcal{B}(\theta) : \\ &\quad \langle \mathbf{z} - \theta, \mathbf{x} - \theta \rangle \leq 1\}. \end{aligned}$$

Then $\mathcal{Q} = \tilde{\mathcal{Q}}$.

We are now ready to prove Theorem 16.

Proof of Theorem 16. Let λ^* be an optimal solution of the dual problem (5) satisfying $A^\top \lambda^* = 0$. This can be chosen because of Lemma 2. Let $\mathcal{Q}(\theta)$, $\tilde{\mathcal{Q}}(\theta)$ and $\mathcal{B}(\theta)$ be as in Lemma 3. Let us define

$$\tilde{\lambda} = \frac{\lambda^*}{\log |\Omega|} + \theta.$$

We will first show that $\tilde{\lambda} \in \tilde{\mathcal{Q}}(\theta)$. We have

$$A^\top \tilde{\lambda} = A^\top \frac{\lambda^*}{\log |\Omega|} + A^\top \theta = A^\top \theta = \mathbf{c}.$$

Thus, $\tilde{\lambda} \in \mathbf{P}_R$. Let $x \in \mathcal{B}$. Then we have

$$\langle \tilde{\lambda} - \theta, \mathbf{x} - \theta \rangle = \frac{\langle \lambda^*, \mathbf{x} - \theta \rangle}{\log |\Omega|} \leq \frac{\log |\Omega|}{\log |\Omega|} = 1$$

where the inequality follows from Lemma 1. Thus $\tilde{\lambda} \in \tilde{\mathcal{Q}}(\theta) = \mathcal{Q}(\theta)$ by Lemma 3. From the definition of $\mathcal{Q}(\theta)$, we have

$$1/\eta \geq \|\tilde{\lambda} - \theta\| = \left\| \frac{\lambda^*}{\log |\Omega|} \right\|.$$

It follows that $\|\lambda^*\| \leq \log |\Omega|/\eta$, finishing the proof. \square

7 PROOF OF THEOREM 11

In this section we prove Theorem 11 by showing how to solve the dual problem (5) using the ellipsoid algorithm.

First, in order to run the ellipsoid algorithm, we need a *first-order oracle*, i.e. we need a procedure to compute $L(\lambda)$ and $\nabla L(\lambda)$. This can be computed by WFOMC using the encoding from Section 2.5. In particular, as discussed in Section 2.5, when both Φ and Φ_0 contain formulas with at most 2 variables, we can compute WFOMC in time polynomial in the size of the domain $|\Delta|$. Hence, in this case we will have a first-order oracle running in time polynomial in $|\Delta|$.

Second, since we have to search for solutions λ^* satisfying $A^\top \lambda^* = 0$, where the matrix A^\top is defined as in Section 6, we need to be able to compute A^\top . For the case when both Φ and Φ_0 contain formulas with at most 2 variables, we can compute the set of vertices

of the relational marginal polytope in time polynomial in $|\Delta|$ as discussed in Section 5. Finding the matrix A^\top is then a straightforward linear algebraic problem. One can then show, using the fact that the number of vertices of the relational marginal polytope is polynomial in $|\Delta|$ and that the representation of these vertices is polynomial in $|\Delta|$ as well, that the number of bits needed to encode A^\top and \mathbf{c} is also polynomial in $|\Delta|$.

Since we have a first-order oracle and we also have means to compute the matrix A^\top and the vector \mathbf{c} which together represent the constraints, we can run the ellipsoid algorithm. However, what remains to be shown is how long the ellipsoid algorithm will need to run in order to obtain a solution with value that is no more than ε from the optimum. We do that next.

Using Theorem 16 and Theorem 1, if we set $R = \log |\Omega|/\eta$ and

$$\beta = -\frac{\varepsilon}{\left(\min_{\lambda \in K, \|\lambda\|_\infty \leq R} L(\lambda) - \max_{\lambda \in K, \|\lambda\|_\infty \leq R} L(\lambda)\right)}$$

then the ellipsoid algorithm will find a solution of the dual problem (5) with value within ε from the optimum in time polynomial in $\log R$, l and $\log(1/\beta)$.

Hence we need to bound β . First, since $L(\lambda) \leq 0$, we can just focus on bounding $\min_{\lambda \in K, \|\lambda\|_\infty \leq R} L(\lambda)$. We have

$$\begin{aligned} -L(\lambda) &= \langle \lambda, \theta \rangle - \log \sum_{\omega \in \Omega} e^{\langle \lambda, Q_\omega(\Phi) \rangle} \\ &\leq |\langle \lambda, \theta \rangle| + \left| \log \sum_{\omega \in \Omega} e^{\langle \lambda, Q_\omega(\Phi) \rangle} \right| \leq l \frac{\log |\Omega|}{\eta} \\ &\quad + \log \left(|\Omega| \cdot \exp \left(l \frac{\log |\Omega|}{\eta} \right) \right) \leq (2l+1) \frac{\log |\Omega|}{\eta}. \end{aligned}$$

Hence, $L(\omega) \geq -(2l+1) \frac{\log |\Omega|}{\eta}$ and $\beta \geq \frac{\varepsilon \eta}{(2l+1) \log |\Omega|}$. It follows that the number of WFOMC calls which the ellipsoid algorithm needs to run is polynomial in $\log(\log |\Omega|/\eta)$, $\log((2l+1) \log |\Omega|/(\varepsilon \eta))$ and l . Finally, noting that each of these calls can be performed in time polynomial in $|\Delta|^c$ and $\log |\Omega|/\eta$ (recall that $\log |\Omega|/\eta$ defines the bounding box where we need to search) and that $\log |\Omega| = O(|\Delta|^{c'})$ finishes the proof (here the constant c depends on Φ and Φ_0 and the constant c' depends on the given first-order language \mathcal{L}). \square

8 PROOF OF THEOREM 12

Here we prove Theorem 12. For that we also need the following lemma, which is just a reformulation of Lemma A.4 from [11] using our notation.

Lemma 4. *Let λ^* be an optimal solution of the dual problem (5) and let λ be such that $L(\lambda) \geq L(\lambda^*) - \varepsilon$.*

Then

$$L(\lambda^*) - L(\lambda) = D_{KL}(p^*||p) \leq \varepsilon$$

where p^* is the MLN given by the formulas from Φ with weights λ^* and p is the MLN given by the same formulas Φ with weights λ .

Next from Pinsker’s inequality we have $\delta_{TV}(p^*, p) \leq \sqrt{D_{KL}(p^*||p)}$ where $\delta_{TV}(p^*, p)$ denotes the total variation distance of p^* and p and p and p^* are as in Lemma 4. Finally, realizing that $|\mathbb{E}_{\omega \sim p^*}[Q_\omega(\Phi)] - \mathbb{E}_{\omega \sim p}[Q_\omega(\Phi)]| \leq \delta_{TV}(p^*, p)$ together with the result in Theorem 11 and with the duality finishes the proof of Theorem 12. \square

9 CONCLUSIONS

We have proved that maximum-likelihood weight learning of MLNs given by formulas with at most 2 variables can be solved in time polynomial in the size of the domain Δ . In order to obtain this result, we framed the learning problem as a relational marginal problem which allowed us to exploit algorithmic techniques from [11]. Some of the new results that we obtained in this paper hold for general MLNs, not just the 2-variable ones. For instance, Theorem 16 holds for all MLNs. The bounds on the number of steps of the ellipsoid algorithm following from the results in Sections 7 and 8 hold for general MLNs as well. We believe that not only the result but also the techniques could be useful for SRL.

We should also stress here that the algorithm described in this paper is meant mostly for theoretical purposes; it is not the most practical one. A more practical algorithm could be obtained if we replaced the ellipsoid algorithm by the projected gradient descent algorithm and designed a more practical variant of the algorithm for construction of relational marginal polytopes.

Acknowledgments A significant part of this work was done while OK was with KU Leuven, supported by Research Foundation - Flanders (project G.0428.15). OK and VK were supported by the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”.

References

- [1] F. Bacchus, A. J. Grove, D. Koller, and J. Y. Halpern. From statistics to beliefs. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 602–608, 1992.
- [2] P. Beame, G. Van den Broeck, E. Gribkoff, and D. Suciú. Symmetric weighted first-order model counting. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 313–328. ACM, 2015.
- [3] B. Bollobás. *Extremal graph theory*. Courier Corporation, 2004.
- [4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [5] R. D. S. Braz, E. Amir, and D. Roth. Lifted first-order probabilistic inference. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1319–1325. Citeseer, 2005.
- [6] D. Buchman and D. Poole. Representing aggregators in relational probabilistic models. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 3489–3495, 2015.
- [7] L. Getoor and B. Taskar. *Introduction to statistical relational learning*, volume 1. MIT press Cambridge, 2007.
- [8] O. Kuželka, Y. Wang, J. Davis, and S. Schockaert. Relational marginal problems: Theory and estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.
- [9] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.
- [10] O. Schulte, H. Khosravi, A. E. Kirkpatrick, T. Gao, and Y. Zhu. Modelling relational statistics with Bayes nets. *Machine Learning*, 94(1):105–125, 2014.
- [11] M. Singh and N. K. Vishnoi. Entropy, optimization and counting. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing (STOC)*, pages 50–59. ACM, 2014.
- [12] G. Van den Broeck, W. Meert, and A. Darwiche. Skolemization for weighted first-order model counting. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 1–10, 2014.
- [13] G. Van den Broeck, N. Taghipour, W. Meert, J. Davis, and L. De Raedt. Lifted probabilistic inference by first-order knowledge compilation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, pages 2178–2185. AAAI Press/International Joint Conferences on Artificial Intelligence, 2011.

- [14] J. Van Haaren, G. Van den Broeck, W. Meert, and J. Davis. Lifted generative learning of markov logic networks. *Machine Learning*, 103(1):27–55, 2016.