Lifted Weight Learning of Markov Logic Networks (Revisited One More Time)

Ondřej Kuželka ONDREJ.KUZELKA@FEL.CVUT.CZ

Vyacheslav Kungurtsev VYACHESLAV.KUNGURTSEV@FEL.CVUT.CZ

Faculty of Electrical Engineering, Czech Technical University in Prague

Yuyi Wang YUWANG@ETHZ.CH

ETH Zurich

Abstract

We revisit the problem of lifted weight learning of Markov logic networks (MLNs). We show that there is an algorithm for maximum-likelihood learning which runs in time polynomial in the size of the domain, whenever the partition function of the given MLN can be computed in polynomial time. This improves on our recent results where we showed the same result with the additional dependency of the runtime on a parameter of the training data, called interiority, which measures how "extreme" the given training data are. In this work we get rid of this dependency. The main new technical ingredient that we exploit are theoretical results obtained recently by Straszak and Vishnoi (Maximum Entropy Distributions: Bit Complexity and Stability, COLT 2019).

1. Introduction

Markov Logic Networks (Richardson and Domingos, 2006) (MLNs) are a prominent statistical relational learning system (Getoor and Taskar, 2007). Generative weight learning of MLNs is typically performed using maximum-likelihood estimation, which is, however, 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 field of *lifted inference* (Braz et al., 2005) and exploited in (Van Haaren et al., 2016) for maximum-likelihood learning of MLNs, where it was shown that gradients of log-likelihood can be computed efficiently. However, this did not provide a bound on the total runtime of the learning algorithm, specifically, because this work was missing a guarantee on the number of iterations of the optimization algorithm. Such a bound on the runtime was later given in (Kuželka and Kungurtsev, 2019; Kuželka and Wang, 2020), mostly building on the results from (Singh and Vishnoi, 2014). Nonetheless, even though the latter runtime bounds are polynomial in the size of the domain for tractable MLNs, they also depend on a parameter that measures how "extreme" given training data are (we discuss this in detail in Section 2.6). This parameter is in general not bounded and can diverge to infinity in some cases. In the present paper, we get rid of the dependency on this parameter by exploiting deep results from (Straszak and Vishnoi, 2019). At least from the theoretical perspective, this seems to be the strongest result that we can hope for. We leave the practical aspects for future work.

2. Background

In this section we describe all the necessary background.

2.1 First Order Logic

We assume a function-free first-order language defined by a set of constants Δ, a set of variables $\mathcal{V}$ and for each $k \in \mathbb{N}$ a set $\mathcal{R}_k$ of k-ary predicates. Variables start with lowercase letters and constants start with uppercase letters. An atom is $r(a_1, ..., a_k)$ with $a_1, ..., a_k \in \Delta \cup \mathcal{V}$ and $r \in \mathcal{R}_k$. A literal is an atom or its negation. A *free* variable is a variable that is not bound by a quantifier. A clause is a universally quantified disjunction of a finite set of literals. A clause in which none of the literals contains any variables is called *ground*. The set of grounding substitutions of a clause α w.r.t. a set of constants Δ is the set $\Theta(\alpha, \Delta) = \{\vartheta_1, ..., \vartheta_m\}$ that contains substitutions to all variables occurring in α using constants from Δ. A possible world ω is represented as a set of ground atoms that are true in ω. The satisfaction relation $\models$ is defined in the usual way: $\omega \models \alpha$ means that the formula α is true in ω. When $\mathbf{x}$ is a list of first-order logic variables then $|\mathbf{x}|$ is used to denote the length of this list.

2.2 Markov Logic Networks

A Markov logic network (MLN, Richardson and Domingos, 2006) 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 Φ induces the probability distribution over 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 groundings of α satisfied in ω, and Z, called *partition function*, is a normalization constant to ensure that p_Φ is a probability distribution.

It is often useful to also allow infinite weights which represent hard logical constraints. For an MLN Φ, let $\Phi_\mathbb{R} = \{(\alpha, w) \in \Phi | w \in \mathbb{R}\}$ be the set of the weighted rules with finite weights and $\Phi_\infty = \{\alpha | (\alpha, +\infty) \in \Phi\}$ be the set of the weighted rules with infinite weights. The distribution given by the MLN Φ is then:

$$p_\Phi(\omega) = \begin{cases} \frac{1}{Z} \exp\left(\sum_{(\alpha,w)\in\Phi} w \cdot N(\alpha, \omega)\right) & \omega \models \Phi_\infty \\ 0 & \text{otherwise} \end{cases}.$$

That is the possible worlds ω that do not satisfy the hard constraints in Φ_∞ have probability zero.

A Note on Notation It is often more convenient to use vector notation. For a list of formulas $\Phi = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ we define $\mathbf{N}(\Phi, \omega) = [N(\alpha_1, \omega), \ldots, N(\alpha_m, \omega)]$. If $\mathbf{w} = [w_1, \ldots, w_m]$ is a vector of weights, we can also write the distribution of an MLN as

$$p_\Phi(\omega) = \frac{1}{Z} \exp\left(\langle \mathbf{w}, \mathbf{N}(\Phi, \omega) \rangle\right)$$

where $\langle ., . \rangle$ denotes the dot product.

2.3 Relational Marginal Problems

Markov logic networks can be seen as solutions to the following maximum entropy problem (Kuželka et al., 2018).[1]

1. For a more general treatment of maximum entropy problems, we refer the reader to (Wainwright et al., 2008).

Given: (i) A list of first-order logic formulas $\Psi = (\alpha_1, \ldots, \alpha_l)$, (ii) Real numbers $\theta = [\theta_1, \ldots, \theta_l]$, (iii) A set of possible worlds Ω over a given domain Δ (Ω is given implicitly by a set of first-order logic sentences that correspond to the hard constraints and by the domain).
Find: A distribution $\{P_\omega : \omega \in \Omega\}$ that is a solution of the following convex optimization problem:

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

$$\sum_{\omega \in \Omega} P_\omega \cdot \mathbf{N}(\Psi, \omega) = \theta \tag{2}$$

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

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 on expected values of the true grounding counts of the formulas $\alpha_1, \ldots, \alpha_l$ and (3) are normalization constraints for the probability distribution.

Assuming there exists a feasible solution satisfying $\forall \omega \in \Omega : P_\omega > 0$ (this will be referred to as the "positivity" assumption), the optimal solution of the above maximum entropy problem is an MLN

$$P_\omega = \frac{1}{Z} \exp\left(\langle \lambda, \mathbf{N}(\Psi, \omega) \rangle\right) \tag{4}$$

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

$$L(\lambda) = \langle \lambda, \theta \rangle - \log \sum_{\omega \in \Omega} e^{\langle \lambda, \mathbf{N}(\Psi, \omega) \rangle} \tag{5}$$

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

2.4 Inference Using Weighted Model Counting

Marginal inference in Markov logic networks, which is also needed for weight learning, can be tackled using *weighted first-order model counting* (Van den Broeck et al., 2011).

Definition 1 (WFOMC, Van den Broeck et al., 2011) *Let $w(P)$ and $\overline{w}(P)$ be functions from predicates to real numbers (we call w and $\overline{w}$* weight functions*) and let Φ be a first-order theory. Then* $\mathrm{WFOMC}(\Phi, w, \overline{w}) = \sum_{\omega \in \Omega : \omega \models \Phi} \prod_{a \in \mathcal{P}(\omega)} w(\textit{Pred}(a)) \prod_{a \in \mathcal{N}(\omega)} \overline{w}(\textit{Pred}(a))$, *where $\mathcal{P}(\omega)$ and $\mathcal{N}(\omega)$ denote the positive literals that are true and false in ω, respectively, and $\textit{Pred}(a)$ denotes the predicate of a (e.g. $\textit{Pred}(\textit{friends}(\textit{Alice}, \textit{Bob})) = \textit{friends}$).*

We now show how to compute the partition function Z of a given MLN using weighted model counting. We proceed as Van den Broeck et al. (2011). 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 Van den Broeck et al. (2011) for the general case). For every weighted formula $(\alpha_i, v_i) \in \Phi$, where $v_i \in \mathbb{R}$ and the free variables in α_i are exactly $x_1, \ldots, x_k$, we create a

new formula $\forall x_1, \ldots, x_k : \xi_i(x_1, \ldots, x_k) \Leftrightarrow \alpha_i(x_1, \ldots, x_k)$ where ξ is a new fresh predicate. Then we set $w(\xi_i) = \exp(v_i)$ and $\overline{w}(\xi_i) = 1$ and for all other predicates we set both w and $\overline{w}$ equal to 1. For every weighted formula $(\alpha_i, +\infty)$, we create a new formula $\forall x_1, \ldots, x_k : \alpha_i(x_1, \ldots, x_k)$. We denote the resulting set of new formulas Γ. It is easy to check that then $WFOMC(\Gamma, w, \overline{w}) = Z$, which is what we needed to compute.

2.4.1 LIFTABILITY

For some classes of first-order logic theories, weighted model counting is a polynomial-time problem. For instance, as shown in (Van den Broeck et al., 2014), 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 polynomial in the number of elements in the domain Δ. This also means that computing the partition function of 2-variable MLNs can be done in time polynomial in the size of the domain. Within statistical relational learning, the term used for problems that have such polynomial-time algorithms is *domain liftability* (Van den Broeck et al., 2011).

Definition 2 (Domain liftability) *An algorithm for computing the partition function Z of an MLN $\Phi = \{(\alpha_1, \lambda_1), \ldots, (\alpha_l, \lambda_l)\}$, where each λ_i is represented by two numbers*[2] *$a_i, b_i \in \mathbb{N}$ as $\lambda_i = \ln a_i - \ln b_i$, is said to be* domain-lifted *if it runs in time polynomial in the size of the domain Δ and in the number of bits needed to encode the numbers a_i and b_i. A class of MLNs is said to be* domain-liftable *if there is a domain lifted algorithm for computing the partition function Z for MLNs from this class.*

The definition that we use here differs slightly from the original definition by Van den Broeck (Van den Broeck et al., 2011) in that it also requires lifted algorithms to depend polynomially on the size of the representation of the formulas' weights. A justification for this alternative definition follows from the work of Jaeger (Jaeger, 2015) (Section 4.2). In particular, all existing domain-lifted inference algorithms are also domain-lifted according to our definition. Another small technical difference is that we define domain-liftability directly in terms of complexity of computing the partition function Z.

2.5 Relational Marginal Polytopes

Here we define relational marginal polytopes (Kuželka et al., 2018), which represent the expected values for the vectors of grounding counts of some given formulas that are possible.[3]

Definition 3 (Relational marginal polytope) *Let Ω be a set of all possible worlds on domain Δ and $\Psi = (\alpha_1, \ldots, \alpha_m)$ be a list of formulas. We define the relational marginal polytope $\mathbf{RMP}(\Psi, \Omega)$ w.r.t. Ψ as $\mathbf{RMP}(\Psi, \Omega) = \{(x_1, \ldots, x_m) \in \mathbb{R}^m : \exists$ dist. on Ω s.t. $\mathbb{E}[N(\alpha_1, \omega)] = x_1 \wedge \cdots \wedge \mathbb{E}[N(\alpha_m, \omega)] = x_m\}$.*

2. The restriction on the representation of the weights ensures that the partition function will always be a rational number. Moreover, one can verify that the number of bits needed to represent the partition function will also be polynomial in the number of bits needed to represent the numbers a_i, b_i and in the domain size $|\Delta|$.
3. In our previous works (Kuželka et al., 2018; Kuželka and Kungurtsev, 2019; Kuželka and Wang, 2020) we worked with relational marginal polytopes which were rescaled versions of the polytopes defined here. The reason was that we were interested in learning MLNs over different domain sizes. Since we are not primarily interested in that in this paper, we opted for the simpler definition, which was called *integer relational marginal polytope* in (Kuželka and Wang, 2020).

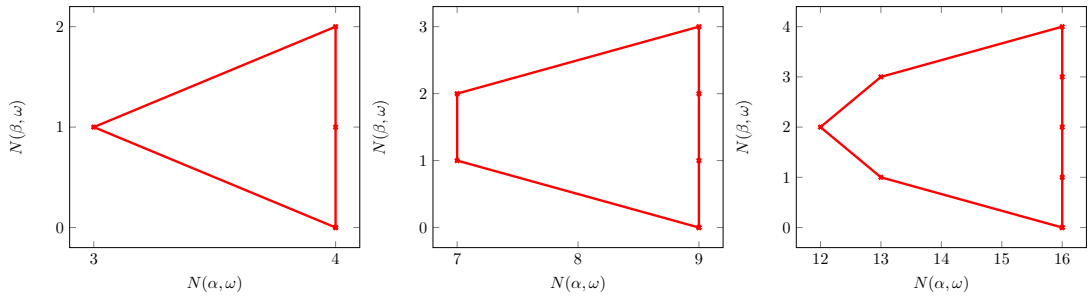

Figure 1: Examples of three relational marginal polytopes given by the first-order logic formulas $\alpha = sm(x) \wedge fr(x,y) \Rightarrow sm(y)$ and $\beta = sm(x)$ for domains of size 2, 3 and 4.

The relational marginal polytope w.r.t. a list of formulas $(\alpha_1, \ldots, \alpha_m)$ can be equivalently defined as the convex hull of the set $\{(N(\alpha_1,\omega), \ldots, N(\alpha_m,\omega)) : \omega \in \Omega\}$.

Example 1 *Let us have formulas $\alpha = sm(x) \wedge fr(x,y) \Rightarrow sm(y)$ and $\beta = sm(x)$ and denote $\Psi = (\alpha, \beta)$. Let Ω_2, Ω_3 and Ω_4 be the sets of all possible worlds over domains of size 2, 3 and 4, respectively, and w.r.t. the first-order language containing only the predicates $sm/1$ and $fr/2$. The three respective relational marginal polytopes $\mathbf{RMP}(\Psi, \Omega_2)$, $\mathbf{RMP}(\Psi, \Omega_3)$ and $\mathbf{RMP}(\Psi, \Omega_4)$ are shown in Figure 1.*

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 $A^{=}\mathbf{x} = \mathbf{c}$ 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 $\{\theta' | A^{=}\theta' = \mathbf{c}, \|\theta' - \theta\|_2 \leq \eta\} \subseteq \mathbf{P}$.*

We need to consider the system of linear equations $A^{=}\mathbf{x} = \mathbf{c}$ in the definition of interiority because the polytope may live in a lower-dimensional subset of the given space.

An important property of polytopes, formalized in (Straszak and Vishnoi, 2019), that we need is their *unary facet complexity*.

Definition 5 (Definition 5.1 in Straszak and Vishnoi, 2019) *Let $P \subseteq \mathbb{R}^m$ be a convex polytope with integer vertices. Let $M \in \mathbb{N}$ be the smallest integer such that P has a description of the form $P = \{x \in \mathbb{R}^m : \langle a_i, x\rangle \leq b_i$, for $i \in I\} \cap H$ where I is a finite index set, $a_i \in \mathbb{Z}^m$, $\|a_i\|_\infty \leq M$ and $b_i \in \mathbb{R}$ for $i \in I$, and H is a linear subspace of $\mathbb{R}^m$. Then we call M the unary facet complexity of P and denote $\mathrm{fc}(P) = M$.*

2.6 Existing Results on Lifted Weight Learning of MLNs

It has been shown in (Kuželka and Kungurtsev, 2019) that maximum likelihood weight learning of Markov logic networks is domain-liftable for the 2-variable fragment of MLNs. This result was then extended to cover all domain-liftable MLNs in (Kuželka and Wang, 2020). Previously, it had been shown in (Van Haaren et al., 2016) that computing the gradients of log-likelihood is domain-liftable for domain-liftable MLNs. What our previous works (Kuželka and Kungurtsev, 2019; Kuželka and

Wang, 2020) added to this was to show that the complete weight learning problem is also domain-liftable, not just the procedure that computes the gradients. This is formally stated in the next theorem.

Theorem 6 *Let $\Psi = (\alpha_1, \ldots, \alpha_l)$ be a list of first-order logic formulas and Φ_0 be a set of first-order logic sentences. Let Ω_{Φ_0} be the set of models of Φ_0 over a given domain Δ. Let $\widehat{\omega} \in \Omega$ be a training example. If computing the partition function of the MLN given by the formulas Ψ on Ω_{Φ_0} is domain liftable, then there is an algorithm which finds weights $\mathbf{w} = (w_1, \ldots, w_l)$ such that the L_1-distance of the distribution of the MLN with these weights and the MLN with optimal weights $\mathbf{w}^*$ maximizing log-likelihood is at most ε. The algorithm runs in time polynomial in $|\Delta|$, $1/\varepsilon$ and $1/\eta$ where η is the interiority of the vector $\mathbf{N}(\Phi, \widehat{\omega})$ in the relational marginal polytope RMP(Φ, Ω_{Φ_0}).*

The polynomial dependency on the interiority parameter $1/\eta$ in the above theorem is problematic as it makes the weight-learning algorithm's runtime dependent not only on the size of the training example but also on its actual structure (i.e., on how "extreme" its statistics are). It is present because we relied on the results about polynomial-time algorithms for maximum entropy problems from (Singh and Vishnoi, 2014). Those results were superseded by the newer results from (Straszak and Vishnoi, 2019) that do not depend on the interiority parameter. In the present paper we build on these newer results and get rid of the interiority parameter for MLN weight learning.

2.7 Polynomial-Time Complexity of Maximum Entropy Problems

Let $\mathcal{F} \subseteq \mathbb{Z}^m$ be a subset of the integer lattice, p be a positive function from $\mathcal{F}$ to $(0; \infty)$ and $\theta \in \mathbb{R}^m$ be a vector. Straszak and Vishnoi (2019) define the following *generalized maximum entropy problem*:[4]

$$\min_{\{q_{\mathbf{n}} \colon \mathbf{n} \in \mathcal{F}\}} \sum_{\mathbf{n} \in \mathcal{F}} q_{\mathbf{n}} \log \frac{q_{\mathbf{n}}}{p(\mathbf{n})} \quad \textit{s.t.} \tag{6}$$

$$\sum_{\mathbf{n} \in \mathcal{F}} q_{\mathbf{n}} \mathbf{n} = \theta \tag{7}$$

$$\forall \mathbf{n} \in \mathcal{F} : q_{\mathbf{n}} \geq 0, \sum_{\mathbf{n} \in \mathcal{F}} q_{\mathbf{n}} = 1 \tag{8}$$

Intuitively when p is a probability distribution (note that it does not have to be), we are asking for a distribution q "closest" to p in KL-divergence that satisfies given marginal constraints (specified by θ).

Assuming θ is in the interior of the marginal polytope, the exact solution of the generalized maximum entropy problem is a probability distribution of the form

$$P(\mathbf{n}) = q_{\mathbf{n}} = \frac{p(\mathbf{n})}{\sum_{\mathbf{n}' \in \mathcal{F}} p(\mathbf{n}') \exp\left(\langle \mathbf{n}', \mathbf{y} \rangle\right)} \exp\left(\langle \mathbf{n}, \mathbf{y} \rangle\right) \tag{9}$$

where $\mathbf{y}$ is the solution of the following problem, which is the Lagrangian dual of the generalized maximum entropy problem:

4. This problem is called *generalized* maximum entropy problem because the classical maximum entropy is its special case when $p(\mathbf{n}) = 1$ or, in general, when $p(\mathbf{n})$ is constant.

$$g(\theta) = \inf_{\mathbf{y}\in\mathbb{R}^m} h(\theta, \mathbf{y}) = \inf_{\mathbf{y}\in\mathbb{R}^m} \log\left(\sum_{\mathbf{n}\in\mathcal{F}} p(\mathbf{n}) \exp\left(\langle \mathbf{n} - \theta, \mathbf{y}\rangle\right)\right). \tag{10}$$

One of the main results of Straszak and Vishnoi (2019) is the structural result given in Theorem 7. This theorem bounds the norm of the vector of parameters $\mathbf{y}$ of an ε-optimal solution of the dual problem (10).

Theorem 7 (Theorem 5.1 in Straszak and Vishnoi, 2019) *Let $\mathcal{F} \subseteq \mathbb{Z}^m$ be a finite subset of the integer lattice and let $d \in [0; +\infty)$ be its diameter, let M be the unary facet complexity of the convex hull P of $\mathcal{F}$. Then, for every function $p : F \to (0; +\infty)$ and for every $\varepsilon > 0$ there exists a number $R > 0$ which is polynomial in m, $\log d$, M, $\max_{\mathbf{n}\in\mathcal{F}} |\log p(\mathbf{n})|$ and $\log(1/\varepsilon)$ such that $\forall\theta \in P, \exists \mathbf{y} \in \mathcal{B}(0, R) : h(\theta, \mathbf{y}) \leq g(\theta) + \varepsilon$, where h and g are as in (10).*

Straszak and Vishnoi then use this theorem to obtain the following important algorithmic result that guarantees polynomial-time solvability of the generalized maximum entropy problem under certain conditions.

Theorem 8 (Theorem 6.1 in Straszak and Vishnoi, 2019) *Let $\mathcal{F} \subseteq \mathbb{Z}^m$ be a finite subset of the integer lattice and let $d \in [0; +\infty)$ be its diameter, let M be the unary facet complexity of the convex hull P of $\mathcal{F}$. Then, there exists an algorithm such that given a probability distribution p on $\mathcal{F}$ (via an evaluation oracle for g_p), $\theta \in P$ and an $\varepsilon > 0$, computes a vector $\mathbf{y} \in \mathbb{R}^m$ with $\|y\| \leq poly(m, M, \log d, \max_{\mathbf{n}\in\mathcal{F}} |\log p(\mathbf{n})|, \log(1/\varepsilon))$ such that $\|q^{\mathbf{y}} - q^*\|_1 \leq \varepsilon$ where q^* is the optimal solution to the generalized maximum entropy problem, $q^{\mathbf{y}}$ is a distribution over $\mathcal{F}$ defined as $q^{\mathbf{y}}_{\mathbf{n}} = \frac{p(\mathbf{n})\exp(\langle \mathbf{n},\mathbf{y}\rangle)}{\sum_{\mathbf{n}'\in\mathcal{F}}\exp(\langle \mathbf{n}',\mathbf{y}\rangle)}$ and g_p is defined as $g_p(\mathbf{x}) = \sum_{\mathbf{n}\in\mathcal{F}} p(\mathbf{n}) \prod_{i=1}^m x_i^{n_i}$ for all $\mathbf{x} \in (0; +\infty)^m$ (where n_i and x_i are the i-th components of $\mathbf{n}$ and $\mathbf{x}$, respectively). The algorithm runs in time polynomial in m, M, $\log d$, $\max_{\mathbf{n}\in\mathcal{F}} |\log p(\mathbf{n})|$, and $\log(1/\varepsilon)$.*

In (Straszak and Vishnoi, 2019), the result from the above theorem is proved using the ellipsoid algorithm (Boyd and Vandenberghe, 2004), which is not the most practical algorithm but has nice theoretical properties.

3. Lifted Weight Learning of Markov Logic Networks Everywhere

In this section we describe our main result which is the following theorem. It shows that one can strengthen the results from (Kuželka and Kungurtsev, 2019; Kuželka and Wang, 2020) and get rid of the dependency on the interiority parameter (cf discussion in Section 2.6).

Theorem 9 *Let $\Psi = (\alpha_1, \ldots, \alpha_l)$ be a list of first-order logic formulas and Φ_0 be a set of first-order logic sentences. Let Ω_{Φ_0} be the set of models of Φ_0 over a given domain Δ. Let $\widehat{\omega} \in \Omega$ be a training example. If computing the partition function of the MLN given by the formulas Ψ on Ω_{Φ_0} is domain liftable, then there is an algorithm which finds weights $\mathbf{w} = (w_1, \ldots, w_l)$ such that the L_1-distance of the distribution of the MLN with these weights and the optimal MLN maximizing log-likelihood is at most ε.[5] The algorithm runs in time polynomial in $|\Delta|$ and $1/\varepsilon$.*

5. When $\mathbf{N}(\Psi, \widehat{\omega})$ lies on the boundary of the respective marginal polytope, we understand the optimal distribution as the solution to the respective relational marginal problem on a subset of the possible worlds Ω (those that can have positive probability in some feasible solution of the relational marginal problem).

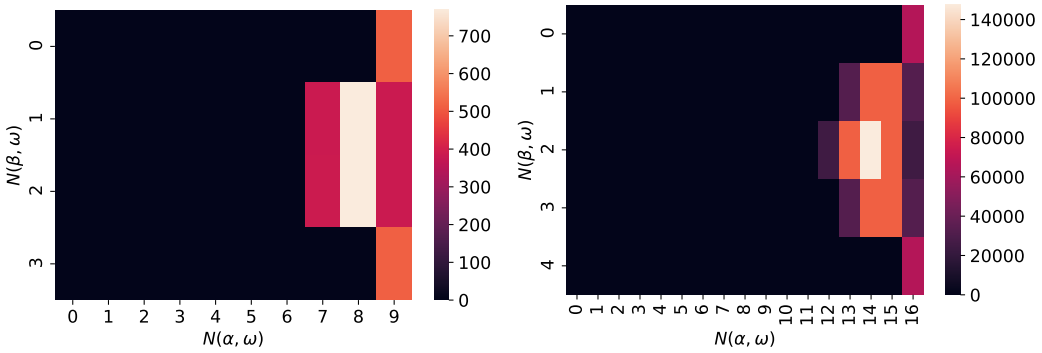

Figure 2: Model-counting functions from Example 2 given by the first-order logic formulas $\alpha = sm(x) \wedge fr(x,y) \Rightarrow sm(y)$ and $\beta = sm(x)$ for domains of size 3 and 4.

Remark 10 *Regarding the meaning of the L_1-distance between two MLNs Φ and Φ^*, note that it holds: if $|\Phi - \Phi^*|_1 \leq \varepsilon$ then $|P_{\omega\sim\Phi}[\omega \models \alpha] - P_{\omega\sim\Phi^*}[\omega \models \alpha]| \leq \frac{\varepsilon}{2}$ (note that this is the same as total variation distance, represented here using first-order logic concepts).*

To prove this theorem, we translate the maximum-likelihood learning problem into the generalized maximum entropy problem and show that the assumptions of Theorem 8 are satisfied for all domain-liftable MLNs. In particular, we need to show the existence of the oracle for g_p and give a polynomial bound on the unary facet complexity of the respective marginal polytopes (which turn out to be the same as relational marginal polytopes, described in Section 2.5).

3.1 The Model-Counting Function

We now describe a useful concept that will be needed in our technical arguments, the *model-counting function* (*MC-function*).

Definition 11 (Model-Counting Function) *Let Ω be a set of possible worlds on a domain Δ and let $\Psi = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a list of first-order logic formulas. We define the model counting function as:* $\mathrm{MC}_{\Psi,\Omega}(\mathbf{n}) = |\{\omega \in \Omega | \mathbf{N}(\Psi, \omega) = \mathbf{n}\}|$.

Intuitively, for any $\mathbf{n} \in \mathbb{Z}^m$, the model counting function gives us the number of possible worlds (from the given set Ω) that satisfy $\mathbf{N}(\Psi, \omega) = \mathbf{n}$.

Example 2 *In Figure 2, we show examples of two MC-functions, $\mathrm{MC}_{\Psi,\Omega}(\mathbf{n})$, for $\Psi = (\alpha, \beta)$, where α and β are as in Example 1, i.e. $\alpha = sm(x) \wedge fr(x,y) \Rightarrow sm(y)$ and $\beta = sm(x)$, for the sets of all possible worlds Ω on domains domains of sizes 3 and 4, respectively.*

3.2 Relational Marginal Problems as Generalized Max-Entropy Problems

We now show how to translate relational marginal problems from Section 2.3 to generalized maximum entropy problems from Section 2.7. This will allow us to exploit the results from (Straszak and Vishnoi, 2019) directly and obtain domain-liftability results from that.

We start by rewriting the relational marginal problem using the model-counting function as follows. First, since we know from the solution of the relational marginal problem (cf Section 2.3) that any two possible worlds ω_1, ω_2 such that $\mathbf{N}(\Psi, \omega_1) = \mathbf{N}(\Psi, \omega_2)$ must have the same probability in the maximum entropy distribution, we can define $P_{\mathbf{n}}$ to be the probability of any possible world ω such that $\mathbf{N}(\Psi, \omega) = \mathbf{n}$. With this we can rewrite the relational marginal problem as:

$$\min_{\{P_{\mathbf{n}} \colon \mathbf{n} \in \mathcal{D}\}} \sum_{\mathbf{n} \in \mathcal{D}} P_{\mathbf{n}} \cdot \mathrm{MC}_{\Psi,\Omega}(\mathbf{n}) \cdot \log P_{\mathbf{n}} \quad \textit{s.t.} \tag{11}$$

$$\sum_{\mathbf{n} \in \mathcal{D}} P_{\mathbf{n}} \cdot \mathrm{MC}_{\Psi,\Omega}(\mathbf{n}) \cdot \mathbf{n} = \theta \tag{12}$$

$$\forall \omega \in \Omega : P_{\mathbf{n}} \geq 0, \sum_{\mathbf{n} \in \mathcal{D}} P_{\mathbf{n}} \cdot \mathrm{MC}_{\Psi,\Omega}(\mathbf{n}) = 1 \tag{13}$$

where $\mathcal{D} = \{0, 1, \ldots, M_1\} \times \{0, 1, \ldots, M_2\} \times \cdots \times \{0, 1, \ldots, M_l\}$ and $M_1 = |\Delta|^{|vars(\alpha_1)|}$, $M_2 = |\Delta|^{|vars(\alpha_2)|}$, $\ldots$, $M_l = |\Delta|^{|vars(\alpha_l)|}$.

Next, we define the set $\mathcal{D}' = \{\mathbf{n} \in \mathcal{D} | \, \mathrm{MC}_{\Psi,\Omega}(\mathbf{n}) \neq 0\}$ and introduce new variables $Y_{\mathbf{n}} \equiv P_{\mathbf{n}} \cdot \mathrm{MC}_{\Psi,\Omega}(\mathbf{n})$ for all $\mathbf{n} \in \mathcal{D}'$, we can rewrite (11), (12), (13) as:

$$\min_{\{Y_{\mathbf{n}} \colon \mathbf{n} \in \mathcal{D}'\}} \sum_{\mathbf{n} \in \mathcal{D}'} Y_{\mathbf{n}} \cdot \log \frac{Y_{\mathbf{n}}}{\mathrm{MC}_{\Psi,\Omega}(\mathbf{n})} \quad \textit{s.t.} \tag{14}$$

$$\sum_{\mathbf{n} \in \mathcal{D}'} Y_{\mathbf{n}} \cdot \mathbf{n} = \theta \tag{15}$$

$$\forall \omega \in \Omega : Y_{\mathbf{n}} \geq 0, \sum_{\mathbf{n} \in \mathcal{D}'} Y_{\mathbf{n}} = 1 \tag{16}$$

This optimization problem already has the form of a generalized maximum entropy problem. That means that we can now use the algorithmic results of Straszak and Vishnoi (2019). Now, supposing we can efficiently solve this generalized maximum-entropy problem, how do we "extract" the MLN from it? The ε-optimal solution of the generalized max-entropy problem is sought in the form of a vector of weights $\mathbf{y}$ that represent the distribution via (cf. Eq 9):

$$Y_{\mathbf{n}} = \frac{\mathrm{MC}_{\Psi,\Omega}(n)}{\sum_{\mathbf{n}' \in \mathcal{D}'} \mathrm{MC}_{\Psi,\Omega}(n) \exp(\langle \mathbf{n}', \mathbf{y} \rangle)} \exp(\langle \mathbf{n}, \mathbf{y} \rangle). \tag{17}$$

Since we have $Y_{\mathbf{n}} = P_{\mathbf{n}} \cdot \mathrm{MC}_{\Psi,\Omega}(\mathbf{n})$ and $P_{\mathbf{n}} = P_\omega$ when $\mathbf{N}(\Psi, \omega) = \mathbf{n}$, we also have $P_\omega = \frac{1}{\sum_{\omega' \in \Omega} \exp(\langle \mathbf{N}(\Psi, \omega'), \mathbf{y} \rangle)} \exp(\langle \mathbf{N}(\Psi, \omega), \mathbf{y} \rangle)$. This is already the MLN we wanted to obtain. Therefore the vector of weights $\mathbf{y}$ that we obtain by solving the generalized maximum entropy problem is also the vector of the weights of the MLN.

Now, let let p and p' denote two distributions on Ω given by P_ω and P'_ω (as above). To link the error that we incur by using the ε-optimal solution obtained by solving the generalized maximum entropy problem, we will need to use the following observation. It holds:

$$\|p - p'\|_1 = \sum_{\omega \in \Omega} |P_\omega - P'_\omega| = \sum_{\mathbf{n} \in \mathcal{D}'} \mathrm{MC}_{\Phi,\Omega}(\mathbf{n}) \cdot |P_{\mathbf{n}} - P'_{\mathbf{n}}| = \sum_{\mathbf{n} \in \mathcal{D}'} |Y_{\mathbf{n}} - Y'_{\mathbf{n}}| = \|q - q'\|_1.$$

This means that we can use the bound on the L_1-distance of the approximate solutions of the generalized maximum entropy problem from Theorem 8 to also bound the L_1-distance of the respective MLNs that we extract from the solution.

Finally, what remains in order to get our domain-liftability results is to show that the conditions from Theorem 8 are satisfied for domain-liftable MLNs. Specifically, we need to show that (i) there is an efficient counting oracle for g_p and (ii) that the unary facet complexity of the marginal polytope, which is the convex hull of $\mathcal{D}'$, is polynomial in the domain size. We do that in Sections 3.3 and 3.4, respectively.

A note about interpretation of the optimization problem. The optimization criterion (14) does not have a direct probabilistic interpretation (since $|\operatorname{MC}_{\Psi,\Omega}(\mathbf{n})|$ is not normalized), however, we can replace (14), without changing the optimal solution as (here the equalities of the $\arg\min$'s are valid only subject to the constraints (16), the argument would not work without using these constraints):

$$\arg \min_{\{Y_{\mathbf{n}}\colon \mathbf{n}\in\mathcal{D}'\}} \sum_{\mathbf{n}\in\mathcal{D}'} Y_{\mathbf{n}} \cdot \log \frac{Y_{\mathbf{n}}}{\operatorname{MC}_{\Psi,\Omega}(\mathbf{n})} = \arg \min_{\{Y_{\mathbf{n}}\colon \mathbf{n}\in\mathcal{D}'\}} \left(\sum_{\mathbf{n}\in\mathcal{D}'} Y_{\mathbf{n}} \cdot \log \frac{Y_{\mathbf{n}}}{\frac{\operatorname{MC}_{\Psi,\Omega}(\mathbf{n})}{|\Omega|}} \right).$$

Now $p^*(\mathbf{n}) = \frac{\operatorname{MC}_{\Psi,\Omega}(\mathbf{n})}{|\Omega|}$ is already a probability distribution. Specifically, $p^*(\mathbf{n})$ is the probability that $N(\Psi,\omega) = \mathbf{n}$ for a possible world ω drawn uniformly from Ω. We can therefore interpret the optimization problem as asking for a distribution q that satisfies the given marginal constraints and is "closest" to the distribution p^* in terms of the KL-divergence $KL(q||p^*)$.

3.3 Counting Oracles

One of the assumptions in Theorem 8 is access to a polynomial-time oracle for the generalized counting function $g_p(\mathbf{x}) = \sum_{\mathbf{n}\in\mathcal{F}} p(\mathbf{n}) \prod_{i=1}^m x_i^{n_i}$ for $\mathbf{x} \in (0,+\infty)^m$. In our case, $p(\mathbf{n}) = \operatorname{MC}_{\Psi,\Omega}(\mathbf{n})$ and $\mathcal{F} = \mathcal{D}'$ (defined in Section 3.2). Now $g_p(\mathbf{x})$ can be also written as:

$$g_p(\mathbf{x}) = \sum_{\mathbf{n}\in\mathcal{D}'} \operatorname{MC}_{\Psi,\Omega}(\mathbf{n}) \cdot \prod_{i=1}^m \exp\left(\ln x_i \cdot n_i\right) = \sum_{\mathbf{n}\in\mathcal{D}'} \operatorname{MC}_{\Psi,\Omega}(\mathbf{n}) \cdot \exp\left(\sum_{i=1}^m \ln x_i \cdot n_i\right) =$$
$$\sum_{\mathbf{n}\in\mathcal{D}'} \operatorname{MC}_{\Psi,\Omega}(\mathbf{n}) \cdot \exp\left(\langle \mathbf{w}, \mathbf{n}\rangle\right) = \sum_{\omega\in\Omega} \exp\left(\langle \mathbf{w}, \mathbf{N}(\Psi,\omega)\rangle\right),$$

where $\mathbf{w} = [\ln x_1, \ldots, \ln x_m]$. The last expression is nothing else than the partition function of an MLN with formulas from Ψ and weights $\mathbf{w} = [\ln x_1, \ldots, \ln x_m]$. So, by definition of domain liftability, we have an oracle for g_p whenever the MLN at hand is domain liftable, which is exactly the result we needed.[6] In particular, note that we do not need to compute the MC-function explicitly at any point!

6. Here we note that the vector $\mathbf{x} = [x_1, \ldots, x_m]$ contains only some finite precision numbers. This is because these numbers are passed to the g_p-oracle from the algorithm from (Straszak and Vishnoi, 2019), which is based on the ellipsoid method that only searches for approximate solutions. For theoretical details about how the analysis of the ellipsoid algorithm deals with finite precision arithmetic we refer to (Grötschel et al., 1988). As a result, the WFOMC oracle which gets the numbers $e^{\ln x_i \cdot n}$ will only have to deal with rational numbers of bounded bit-lengths (we note that we talked about the representation issues when defining domain-liftability in Section 2.4.1 precisely for this reason).

3.4 Unary Facet Complexity of Relational Marginal Polytopes

The method from Theorem 8 runs in time polynomial in the unary facet complexity of the marginal polytope. The marginal polytope is the convex hull of the set $\mathcal{D}'$ which, as can be verified straightforwardly, is the relational marginal polytope of Ψ over the domain Δ (cf. Section 2.5). The next proposition shows that the unary facet complexity of this polytope is polynomial in the domain size.

Proposition 12 *The unary facet complexity of every relational marginal polytope is polynomially bounded (in the size of the domain)* Δ.

Proof To prove this lemma, we only need to show that, the coefficients of the inequalities describing any relational marginal polytope are polynomially bounded. First, we observe that every entry of every integer point in a relational marginal polytope is $O(m^k)$ where m is the domain size. A facet of the polytope is a hyperplane or an intersection of at most $d = |\Psi|$ hyperplanes each of which passes through d integer points $x_1, x_2, \ldots, x_d$. The normal vector of the hyperplane is a vector orthogonal to all the $(x_1 - x_d), \ldots, (x_{d-1} - x_d)$. Note that the entries in $(x_i - x_d)$ are still $O(m^k)$ for all i. Let M be the matrix whose i-th row is $(x_i - x_d)$. The i-th coefficient of the hyper-plane-defining equation is the determinant of M_i where M_i is M removing i-th column. This determinant is at most $O(d!m^{dk})$ (so still poly in m) and is an integer. ■

The algorithm from (Straszak and Vishnoi, 2019) actually needs an explicit bound on the unary facet complexity of the marginal polytope. While we could also use a crude upper bound based on the reasoning from the proof of the above proposition, we can also proceed as follows. For any domain-liftable MLN, to compute the unary facet complexity, we can construct the respective relational marginal polytope in polynomial time using a WFOMC oracle as shown in (Kuželka and Wang, 2020). Once we have the marginal polytope, computing a polynomial bound on its unary facet complexity is straightforward.

3.5 Finishing the Proof of Theorem 9

We are practically done. In Section 3.2 we have shown how to convert a relational marginal problem to a generalized maximum entropy problem and how to extract the solution to the original problem from it. The dual of the relational marginal problem is the maximum likelihood problem, assuming the domain we want to model is of the same size as that of the training example (Kuželka et al., 2018). Hence, if we can show that we can solve the resulting generalized maximum entropy problem efficiently, the result that we need to prove will follow. To do that we need to show that the assumptions of Theorem 8 are satisfied. That means that we need to provide an efficient oracle for the generalized counting function g_p, which we did in Section 3.3, and to show that the unary facet complexity is bounded by a polynomial, which we did in Section 3.4. So we are done.

4. Conclusions

In this work we improved our results from (Kuželka and Kungurtsev, 2019; Kuželka and Wang, 2020). Specifically, we removed the dependency on the interiority parameter $1/\eta$ from the runtime of the MLN weight learning algorithm by exploiting the results from (Straszak and Vishnoi, 2019). The argument in the present paper is also, arguably, simpler than the argument used in our previous

works, since here we reduce the problem directly to the generalized maximum entropy problem studied by Straszak and Vishnoi.

References

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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.

L. Getoor and B. Taskar. *Introduction to statistical relational learning*, volume 1. MIT press Cambridge, 2007.

M. Grötschel, L. Lovász, and A. Schrijver. *The Ellipsoid Method*, pages 64–101. Springer Berlin Heidelberg, 1988. ISBN 978-3-642-97881-4.

M. Jaeger. Lower complexity bounds for lifted inference. *TPLP*, 15(2):246–263, 2015.

O. Kuželka and V. Kungurtsev. Lifted weight learning of markov logic networks revisited. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 1753–1761, 2019.

O. Kuželka and Y. Wang. Domain-liftability of relational marginal polytopes. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, pages –, 2020.

O. Kuželka, Y. Wang, J. Davis, and S. Schockaert. Relational marginal problems: Theory and estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

M. Richardson and P. Domingos. Markov logic networks. *Mach Learn*, 62(1-2):107–136, 2006.

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.

D. Straszak and N. K. Vishnoi. Maximum entropy distributions: Bit complexity and stability. In *Conference on Learning Theory, COLT*, pages 2861–2891, 2019.

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.

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.

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.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.