Relational Marginal Problems: Theory and Estimation

Ondřej Kuželka
Cardiff University, UK
KuzelkaO@cardiff.ac.uk

Yuyi Wang
ETH Zurich, Switzerland
yuwang@ethz.ch

Jesse Davis
KU Leuven, Belgium
jesse.davis@cs.kuleuven.be

Steven Schockaert
Cardiff University, UK
SchockaertS1@cardiff.ac.uk

Abstract

In the propositional setting, the *marginal problem* is to find a (maximum-entropy) distribution that has some given marginals. We study this problem in a relational setting and make the following contributions. First, we compare two different notions of relational marginals. Second, we show a duality between the resulting relational marginal problems and the maximum likelihood estimation of the parameters of relational models, which generalizes a well-known duality from the propositional setting. Third, by exploiting the relational marginal formulation, we present a statistically sound method to learn the parameters of relational models that will be applied in settings where the number of constants differs between the training and test data. Furthermore, based on a relational generalization of *marginal polytopes*, we characterize cases where the standard estimators based on feature's number of true groundings needs to be adjusted and we quantitatively characterize the consequences of these adjustments. Fourth, we prove bounds on expected errors of the estimated parameters, which allows us to lower-bound, among other things, the effective sample size of relational training data.

Introduction

Statistical Relational Learning (SRL, Getoor and Taskar, eds., 2007) is concerned with learning probabilistic models from relational data. Many popular SRL frameworks, such as Markov Logic Networks (MLNs, Richardson and Domingos 2006), use weighted logical formulas to encode statistical regularities that hold for the considered problem. Typically, the maximum (pseudo-)likelihood weights of the formulas are estimated from training data, which is usually a single large example (e.g. a social network). This is problematic for two reasons. First, the weights that are learned from this single training example are in general not optimal for examples of different sizes (Jain, Kirchlechner, and Beetz 2007). This turns out to be a fundamental problem, which cannot simply be solved by rescaling the weights (Shalizi and Rinaldo 2013). Second, without making further assumptions, it is difficult to provide any statistical guarantees about the learned weights.

In this paper, we approach parameter estimation in SRL from a novel direction, by introducing the notion of a *rela-tional marginal problem*. In the propositional case (Wainwright and Jordan 2008), marginal problems entail finding a maximum-entropy distribution which has the given marginal probabilities. A well-known property of such problems is that they are dual to the maximum-likelihood estimation of the parameters of an undirected graphical model (where "dual" is in the sense of convex optimization). In relational marginal problems, we are similarly looking for a maximum-entropy distribution which satisfies some given statistics – *relational marginals*. However, we also need to define what these relational marginals are. Thus, first, we describe two different types of *relational marginals*, which differ in the kinds of statistics that are provided.[1] The first type is based on relational marginal distributions (Kuželka, Davis, and Schockaert 2017) and the second is based on Halpern-style random substitution semantics (Bacchus et al. 1992). Second, for both types of statistics, we establish a relational counterpart of the duality between maximum-likelihood estimation and max-entropy marginal problems. Interestingly, for the latter model, the corresponding dual is MLNs.

Third, the relational marginal perspective allows us to learn parameters for domains that have different sizes (i.e., number of constants) than the training data. The basic idea to achieve this is simple. We assume that the training data is a sample of the data that we want to model. For example, imagine trying to model all of Facebook based on a sampled subset of Facebook users along with all relations among them. Assuming the sample is a large enough and was obtained in a suitable way (which is not always the case in practice – we discuss this issue later), the parameters of the marginals estimated from the sample should be close to the respective parameters for the whole network. Then, instead of using a model learned by optimizing the likelihood on the training data, we use a model obtained as a solution of the corresponding relational marginal problem with a domain of the required size. We may end up with estimated parameters for which the relational marginal problem has no solution. Therefore, we propose a method for adjusting the estimated parameters that enables a solution and characterize its effect on the estimates. Then we also introduce the *relational*

---

[1] All statistics in this paper are based on universally quantified formulas; we do not consider formulas with existential quantifiers.

marginal polytopes, which allows us to provide conditions under which the unbiased unadjusted estimate will be valid ("realizable") for domains of any size.

In addition, the relational marginal view of the parameter learning problem, can be thought of as consisting of two decoupled problems: estimation of the parameters of marginals and optimization to obtain the max-entropy distributions. Thus, to better understand parameter learning from relational data, it is important to characterize how accurate the estimates are. Assuming that all subsamples of the data being modeled are sampled with the same probability, we derive bounds on expected error, that is, the expected difference between the parameters obtained from the subsample and parameters that could be theoretically computed if the whole dataset were accessible (e.g. the whole Facebook). From this, we can also obtain lower-bounds on the effective sample size for relational data.

The paper is structured around addressing the following four questions about relational marginal problems:

1. What should the relational marginals be? (Section *Two Types of Relational Marginals*)

2. What are the max-entropy distributions with the given relational marginals, and how can we find them? (Section *Max-Entropy Models*)

3. When are relational marginal problems realizable, and how can we adjust them when they are not? How can we adjust learning to account for differences in domain sizes? (Section *Realizability*)

4. How accurate are the parameter estimates of relational marginals, and what are the links with realizability? (Sections *Relational Marginal Polytopes* and *Estimation*)

Proofs of all propositions stated in the main text and details of the duality derivations are contained in the online appendix.[2]

Preliminaries

This paper considers a function-free first-order logic language $\mathcal{L}$, which is 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 and use the domain size to refer to the $|Const|$. For $a_1, ..., a_k \in Const \cup Var$ and $R \in Rel_k$, we call $R(a_1, ..., a_k)$ an atom. If $a_1, .., a_k \in Const$, 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 possible world ω is defined as a set of ground atoms. The satisfaction relation $\models$ is defined in the usual way.

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 term. As is commonly done in statistical relational learning, we use the unique names assumption, meaning that $c_1 \neq c_2$ whenever c_1 and c_2 are different constants. A first-order universally quantified formula

---

[2]https://arxiv.org/abs/1709.05825

α is said to be *proper* if $\alpha\vartheta$ is trivially true whenever ϑ is not injective. For instance, the formula $\forall X, Y : fr(X, Y)$ is not proper whereas the formula $\forall X, Y : fr(X, Y) \lor X = Y$ is proper. We sometimes omit the universal quantifiers and simply write, e.g. $fr(X, Y) \lor X = Y$.

A (global) example is a pair $(\mathcal{A}, \mathcal{C})$, with $\mathcal{C}$ a set of constants and $\mathcal{A}$ a set of ground atoms which only use constants from $\mathcal{C}$. Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example and $\mathcal{S} \subseteq \mathcal{C}$. The fragment $\Upsilon\langle S \rangle = (\mathcal{B}, \mathcal{S})$ is defined as the restriction of Υ to the constants in $\mathcal{S}$, i.e. $\mathcal{B}$ is the set of all atoms from $\mathcal{A}$ which only contain constants from $\mathcal{S}$. Two examples $\Upsilon_1 = (\mathcal{A}_1, \mathcal{C}_1)$ and $\Upsilon_2 = (\mathcal{A}_2, \mathcal{C}_2)$ are isomorphic, denoted as $\Upsilon_1 \approx \Upsilon_2$, if there exists a bijection $\sigma : \mathcal{C}_1 \rightarrow \mathcal{C}_2$ such that $\sigma(\mathcal{A}_1) = \mathcal{A}_2$, where σ is extended to ground atoms in the usual way. When $\mathcal{C}$ is a set of constants and Φ_0 a set of closed formulas, $\Pi(\mathcal{C}, \Phi_0)$ denotes the set of all $\Upsilon = (\mathcal{A}, \mathcal{C})$ such that $\Upsilon \models \Phi_0$ (we can think of Φ_0 as a set of constraints).

A Markov logic network (MLN, Richardson and Domingos 2006) is a set of weighted formulas (α, w), $w \in \mathbb{R}$ and α 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 $\mathcal{C}$, called the domain. An MLN is seen as a template that defines a Markov random field. Specifically, an MLN $\mathcal{M}$ induces the following probability distribution on the set of global examples[3] Υ:

$$p_{\mathcal{M}}(\Upsilon) = \frac{1}{Z} \exp \left(\sum_{(\alpha, w) \in \mathcal{M}} w \cdot n(\alpha, \Upsilon) \right) \qquad (1)$$

where $n(\alpha, \Upsilon)$ is the number of groundings of α that are satisfied in Υ, and Z is a normalization constant to ensure that $p_{\mathcal{M}}$ is a probability distribution.

Weights of MLNs are typically estimated using maximum (pseudo)likelihood from a given global example. When the MLN with weights learned in this way is used as a probabilistic model with a domain of different size, there are no guarantees regarding the induced distribution. This is most obvious when the MLN contains formulas having different numbers of variables. Then, keeping the weights fixed, the formulas with the highest number of variables often completely dominate the others if we increase the domain size. While some simple cases could be solved by normalizing the counts $n(\alpha, \Upsilon)$, in general this is not the case. Shalizi and Rinaldo (2013) list the example of modelling homophily in social networks; we refer to their paper for details.

Two Types of Relational Marginals

Typically, parameters for a statistical relational model are estimated from a single example of a relational structure that consists of a large set of ground atoms $\mathcal{A}$. Intuitively, the goal is to learn a probability distribution of such relational structures. The challenge is how to estimate the distribution from a single example. One solution is based on the assumption that the relational structure has repeated regularities. Then, statistics about these regularities can be computed for

---

[3]What we call a global example in this paper is usually called a *possible world* in the MLN literature.

small substructures of the train example and used to construct a distribution over large relational structures. Thus, the next issue is how to construct the fragments and compute statistics on them. Next, we discuss two possible ways to do so, which we will refer to as Model A and Model B.

Model A

The first approach to constructing fragments is from (Kuželka, Davis, and Schockaert 2017). It repeatedly samples subsets $\mathcal{S} \subseteq \mathcal{C}$ of the constants from the given example $\Upsilon = (\mathcal{A}, \mathcal{C})$ and then builds one training example $\Upsilon\langle\mathcal{S}\rangle$ for each $\mathcal{S}$. However, the training examples must consider isomorphic classes of constants to account for the fact that each fragment will contain different constants. This is accomplished by using the notion of *local examples*.

Definition 1 (Local example). *Let $k \in \mathbb{N}$. A local example of width k is a pair $\omega = (\mathcal{A}, \{1, ..., k\})$, where $\mathcal{A}$ is a set of ground atoms that contain only constants from the set $\{1, 2, \ldots, k\}$. For an example $\Upsilon = (\mathcal{A}, \mathcal{C})$ and $\mathcal{S} \subseteq \mathcal{C}$, we write $\Upsilon[\mathcal{S}]$ for the set of all local examples of width $|\mathcal{S}|$ which are isomorphic to $\Upsilon\langle\mathcal{S}\rangle$.*

To distinguish local examples from global examples, we will denote them using lower case Greek letters such as ω instead of upper case letters such as Υ.

Example 1. *Let $\Upsilon = (\{fr(alice, bob), fr(bob, alice), fr(bob, eve), fr(eve, bob), sm(alice)\}, \{alice, bob, eve\})$, i.e. the only smoker is alice and the friendship structure is:*

Then $\Upsilon\langle\{alice, bob\}\rangle = (\{sm(alice), fr(alice, bob), fr(bob, alice)\}, \{alice, bob\})$, $\Upsilon[\{alice, bob\}] = \{(\{fr(1, 2), fr(2, 1), sm(1)\}, \{1, 2\}), (\{fr(2, 1), fr(1, 2), sm(2)\}, \{1, 2\})\}$.

This leads to a natural definition of a probability distribution over local examples of width k.

Definition 2 (Relational marginal distribution of a global example). *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example and $k \in \mathbb{N}$. The relational marginal distribution of Υ of width k is a distribution $P_{\Upsilon,k}$ over local examples, where $P_{\Upsilon,k}(\omega)$ is defined as the probability that ω is sampled by the following process: (i) Uniformly sample a subset $\mathcal{S}$ of k constants from $\mathcal{C}$. (ii) Uniformly sample a local example ω from the set $\Upsilon[\mathcal{S}]$. For a closed formula α without constants, we also define, its probability: $P_{\Upsilon,k}(\alpha) = \sum_{\omega:\omega\models\alpha} P_{\Upsilon,k}(\omega)$.*

We will call a pair (α, p), where α is a constant-free closed formula and $p \in [0; 1]$, a *relational marginal constraint*. We may also interpret the probability $P_{\Upsilon,k}(\alpha)$ of a closed constant-free formula α as the probability that α is true in a restriction $\Upsilon\langle\mathcal{S}\rangle$ of Υ to a randomly sampled subset $\mathcal{S}$ of k constants from Υ. Thus, if we are only interested in the probabilities of closed constant-free formulas, we do not have to refer to local examples. Local examples are important because relational marginal distributions defined using them are themselves probability distributions on possible worlds, which is both nice conceptually and convenient (as it means that we can model relational marginals of Model A using

any standard propositional probabilistic model). Arguably, this property also makes Model A more interpretable and hence easier to explain to non-specialists.

Global examples may also be assumed to be sampled from some distribution and we define the corresponding marginal distributions induced by such distributions accordingly. When $P(\Upsilon)$ is a distribution over finite global examples from a possibly countably infinite set Ω, then the marginal distribution of width k is a distribution P_k over local examples where $P_k(\omega)$ is defined as $P_k(\omega) = \sum_{\Upsilon\in\Omega} P(\Upsilon) \cdot P_{\Upsilon,k}(\omega)$. For a closed formula α without constants, we also analogically define: $P_k(\alpha) = \sum_{\omega:\omega\models\alpha} P_k(\omega)$. In other words, a relational marginal distribution is a mixture of (possibly countably many) relational marginal distributions of global examples.

Proposition 2. *Let $P(\Upsilon)$ be a distribution on domain size n and $k \leq n$ be an integer. Let $\Omega_\alpha = \{\Upsilon\langle\mathcal{S}\rangle : \mathcal{S} \subseteq \mathcal{C}, |\mathcal{S}| = k, \Upsilon\langle\mathcal{S}\rangle \models \alpha\}$ where $\Upsilon = (\mathcal{A}, \mathcal{C})$ is sampled according to the distribution $P(\Upsilon)$. Then, for a closed constant-free formula α, $\widehat{p}_\alpha = |\Omega_\alpha| \cdot \begin{pmatrix} n \\ k \end{pmatrix}^{-1}$ is an unbiased estimate of $P_k(\alpha)$ for Model A.*

Model B

The second approach is to consider random substitutions, which is in the spirit of existing works (Bacchus et al. 1992; Schulte et al. 2014). Here, the statistics that we collect about Υ are defined as follows.

Definition 3 (Probability of formulas under Model B). *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example and α be a universally quantified formula. Let P_ϑ be a uniform distribution on injective substitutions from the set $\Theta_\alpha = \{\vartheta | \vartheta : vars(\alpha) \to \mathcal{C} \text{ and } \vartheta \text{ is injective}\}$. Then the probability $Q(\alpha)$ of the formula α under model B is defined as*

$$Q_\Upsilon(\alpha) = \sum_{\vartheta\in\Theta_\alpha} \mathbb{1}(\Upsilon \models \alpha\vartheta)P_\vartheta(\vartheta) = \frac{1}{|\Theta|}\sum_{\vartheta\in\Theta_\alpha} \mathbb{1}(\Upsilon \models \alpha\vartheta).$$

Just like for Model A, we extend the definition of the probability of formulas straightforwardly to the case where Υ is not fixed but sampled from some distribution over a countable set Ω: $Q(\alpha) = \sum_{\Upsilon\in\Omega} Q_\Upsilon(\alpha) \cdot P(\Upsilon)$.

Example 3. *Let Υ be as in Example 1. Let*

$$\alpha = \forall X, Y : \neg fr(X, Y) \vee sm(Y),$$
$$\beta = \forall X, Y : \neg fr(X, Y) \vee sm(X) \vee sm(Y).$$

Assume that the relation $fr(., .)$, "friends", is symmetric. Then, the formula α is classically true if all people who are friends with someone, smoke. The formula β is classically true if for every pair of people A and B who are friends, at least one of them smokes. Computing the respective probabilities, we get $P_{\Upsilon,2}(\alpha) = \frac{1}{3}$, $P_{\Upsilon,2}(\beta) = \frac{2}{3}$, $Q_\Upsilon(\alpha) = \frac{1}{2}$, $Q_\Upsilon(\beta) = \frac{2}{3}$, which illustrates that in general the "marginal" probabilities given by the two models will differ. The first model might be slightly easier to interpret as the marginal probabilities $P(\gamma)$ correspond to the fraction of the width-k fragments of Υ in which γ is true as a classical logic formula.

We note that the straightforward analogue of Proposition 2 also holds for Model B.

Max-Entropy Models

In this section we show how to compute *models* of given relational marginals under Model A and Model B.

Definition 4 (Model of relational marginals). *Let us have a set of pairs* $\Phi = \{(\alpha_1, \theta_1), \ldots, (\alpha_h, \theta_h)\}$ *of relational marginals, where* α_i *is a closed formula and* $\theta_i \in [0;1]$. *Let* Φ_0 *be a set of formulas, called hard rules (i.e. formulas that are supposed to always hold). We say that a probability distribution* $P(\Upsilon)$ *over worlds satisfying the hard rules from* Φ_0 *is a width-k model of* Φ *iff* $P_k(\alpha_i) = \theta_i$ ($Q_k(\alpha_i) = \theta_i$, *respectively) for all* $(\alpha_i, \theta_i) \in \Phi$.

We will use standard duality arguments from convex optimization (Boyd and Vandenberghe 2004), essentially following (Singh and Vishnoi 2014). For both models, the result will in both cases an exponential-family model and the one for Model B will turn out to be equivalent to MLNs.

Model A

Let $\mathcal{C}$ be a set of constants, $\mathcal{A}$ be the set of all atoms over $\mathcal{C}$ based on some given first-order language and let $\mathcal{C}_k$ denote the set of all k-element subsets of $\mathcal{C}$. Next, let Φ be a set of relational marginals and Φ_0 be a set of hard constraints, i.e. formulas α such that if $\Upsilon \not\models \alpha$ then $P(\Upsilon) = 0$. We assume that there exists at least one distribution P' which is a model of Φ and which satisfies the following *positivity* condition: $P'(\Upsilon) > 0$ for all Υ satisfying the hard constraints (i.e. those for which $\Upsilon \models \Phi_0$).

$$\sup_{\{P_\Upsilon : \Upsilon \in \Pi(\mathcal{C}, \Phi_0)\}} \sum_{\Upsilon \in \Pi_n(\mathcal{C}, \Phi_0)} P_\Upsilon \log \frac{1}{P_\Upsilon} \quad s.t. \quad (2)$$

$$\forall (\alpha_i, \theta_i) \in \Phi :$$

$$\frac{1}{|\mathcal{C}_k|} \sum_{\mathcal{S} \in \mathcal{C}_k} \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} \mathbb{1}(\Upsilon \langle \mathcal{S} \rangle \models \alpha_i) \cdot P_\Upsilon = \theta_i \quad (3)$$

$$\forall \Upsilon \in \Pi_n(\mathcal{C}, \Phi_0) : P_\Upsilon \geq 0, \quad \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} P_\Upsilon = 1 \quad (4)$$

Next we define $\#_k(\alpha, \Upsilon) = |\{\mathcal{S} \in \mathcal{C}_k : \Upsilon \langle \mathcal{S} \rangle \models \alpha\}|$ as the number of sets $\mathcal{S} \in \mathcal{C}_k$ such that the formula α is classically true in the restriction of Υ to constants in $\mathcal{S}$. In the appendix, we show that the distribution, if it exists, that is a solution to the optimization problem has the following form:

$$P_\Upsilon = \frac{1}{Z} \exp \left(\sum_{\alpha_i \in \Phi} \lambda_i \frac{\#_k(\alpha_i, \Upsilon)}{|\mathcal{C}_k|} \right). \quad (5)$$

The Lagrangian dual problem of the maximum entropy problem is to maximize (where $\lambda_i \in \mathbb{R}$):

$$\sum_{\alpha_i \in \Phi} \lambda_i \theta_i \quad - \quad \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} \lambda_i \frac{\#_k(\alpha_i, \Upsilon)}{|\mathcal{C}_k|}}. \quad (6)$$

Due to the positivity assumption, Slater's condition (Boyd and Vandenberghe 2004) is satisfied and strong duality holds. Consequently, instead of solving the original problem, which has an intractable number of constraints and variables (one variable for each world $\Upsilon \in \Pi(\mathcal{C}, \Phi_0)$), we can solve the dual problem, which only has $|\Phi|$ variables. On the other hand, the optimization criterion of the dual problem may still be computationally hard to solve as it requires weighted counting over worlds in $\Pi(\mathcal{C}, \Phi_0)$. However, in many restricted, but non-trivial, cases, we can exploit lifted weighted model counting techniques in the same way as they were used for maximum-likelihood estimation in Van Haaren et al. (2016).

Let us perform a change of variables $w_i := \lambda_i / |\mathcal{C}_{k_i}|$. This gives us

$$P(\Upsilon) = \frac{1}{Z} \exp \left(\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon) \right) \quad (7)$$

for the distribution and

$$\sum_{\alpha_i \in \Phi} w_i \theta_i |\mathcal{C}_{k_i}| - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon)} \quad (8)$$

for the optimization criterion of the dual problem. Assuming that the marginals were estimated from a global example $\widehat{\Upsilon} \in \Pi(\mathcal{C}, \Phi_0)$ (note that here the domain $\mathcal{C}$ is the same as the domain of the global examples over which the distribution is defined) and that they still satisfy the positivity assumption, we can also rewrite (8) as

$$\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \widehat{\Upsilon}) - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon)}.$$

It is straightforward to check that this is the same as directly optimizing the log-likelihood of $\widehat{\Upsilon}$. Thus, this is a relational analogue of the well-known duality of maximum likelihood and maximum entropy (Wainwright and Jordan 2008). Note that it is important for the duality of maximum likelihood and maximum entropy that both the $\widehat{\Upsilon}$, from which we estimated the parameters, and the global examples over which the distribution is computed have the same domain size. The Section *Realizability* will address cases where the domain sizes of the training and testing data differ.

Model B

Like for Model A, we can construct a convex optimization problem to obtain a maximum-entropy distribution with the given relational marginals under Model B. This problem's optimization criterion is the same as (2). To obtain the constraints enforcing the marginals, we can replace the summation over subsets of constants in $\mathcal{C}$ in Equation (3) by a summation over substitutions from Θ_{α_i}, where Θ_{α_i} is defined as in Definition 3. This yields the following set of constraints for all $(\alpha_i, \theta_i) \in \Phi$:

$$\sum_{\vartheta \in \Theta_{\alpha_i}} \frac{1}{|\Theta_{\alpha_i}|} \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} \mathbb{1}(\Upsilon \models \alpha_i \vartheta) \cdot P_\Upsilon = \theta_i. \quad (9)$$

Using basically the same reasoning as for Model A, we arrive at the following form of the probability distribution

$$P(\Upsilon) = \frac{1}{Z} \exp\left(\sum_{\alpha_i \in \Phi} w_i \cdot n(\alpha_i, \Upsilon)\right) \qquad (10)$$

where $n(\alpha_i, \Upsilon)$ is the number of groundings $\alpha_i \vartheta$ of the formula α_i, where all ϑ's are injective, which are true in Υ. This distribution is identical to the one for MLNs which only contain proper formulas (because of the injectivity requirement in the definition of Model B). The only difference with the distribution of Model A is the use of $n(\alpha_i, \Upsilon)$ instead of $\#_k(\alpha_i, \Upsilon)$. The dual optimization criterion for Model B then becomes to maximize

$$\sum_{\alpha_i \in \Phi} w_i \theta_i |\Theta_{\alpha_i}| - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot n(\alpha_i, \Upsilon)} \quad (11)$$

which can be rewritten, when θ_i's are estimated from some $\widehat{\Upsilon} = (\widehat{\mathcal{A}}, \widehat{\mathcal{C}})$ with $|\widehat{\mathcal{C}}| = |\widehat{\mathcal{C}}|$, as $\sum_{\alpha_i \in \Phi} w_i \cdot n(\alpha_i, \widehat{\Upsilon}) - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot n(\alpha_i, \Upsilon)}$. This is the same as log-likelihood of $\widehat{\Upsilon}$ w.r.t. the MLN given by (11), assuming that all the formulas in the MLN are proper. Thus, when the size of the domain of the training example $\widehat{\Upsilon}$ is the same as the cardinality of the domain of the modeled distribution, the max-entropy relational marginal problem in Model B is the same as maximum-likelihood estimation in MLNs.

Realizability

Not all relational marginals have a model for a given domain size (or even any model at all). This is a problem if we want to estimate relational marginals on some given global example $\widehat{\Upsilon}$ and then use them to obtain a distribution on global examples that have a different domain size (which we address in Section *Estimation*). The duality between maximum-likelihood estimation and max-entropy relational marginal problems discussed in the previous section only holds when the training data and the distribution that we want to model have the same domain size.

In this section, we will show how to obtain relational marginals for any domain size. Accomplishing this requires replacing the consistency of relational marginal estimators with a weaker notion. However, we can still bound the difference between the unbiased and the adjusted estimates. More importantly, the difference will tend to zero as the domain size of the examples from which the respective marginals were estimated increases.

First, it is easy to see that if a model exists for a given set of marginals and domain size, then there is also such a model for smaller domain sizes, as the next proposition asserts.[4]

Proposition 4. *For Model A, if there is a width-k model of Φ on domain of size n then there is also a width-k model of Φ on domain of size m for any m satisfying $n \geq m \geq k$. For Model B, if there is a model of Φ on domain of size n*

---

[4]Proposition 4 is one of the results which would not hold for Model B if we did not require injectivity of the randomly sampled substitutions in the definition of Model B.

then there is also a model of Φ on domain of size m for any m satisfying $n \geq m \geq \max_{\alpha \in \Phi} |vars(\alpha)|$.

Next, we give an example of a relational marginal distributions that does not have a model for arbitrary domain cardinalities.

Example 5. *Let $\mathcal{L}$ consist of predicate symbols $e/2, r/1, g/1, b/1$ and $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example where*

$$\mathcal{A} = \{e(v_1, v_2), e(v_2, v_1), e(v_2, v_3), e(v_3, v_2), e(v_1, v_3),$$
$$e(v_3, v_1), r(v_1), g(v_2), b(v_3)\}, \quad \mathcal{C} = \{v_1, v_2, v_3\}.$$

Let $k = 2$ be the width of local examples. And let $F(X_1, X_2) \stackrel{\text{def}}{=} X_1 \neq X_2 \wedge \neg e(X_1, X_1) \wedge e(X_1, X_2) \wedge e(X_2, X_1) \wedge \neg e(X_2, X_2)$. Then we can estimate, for instance, the following marginals from Υ under Model A:

$$P[\exists X_1, X_2 : F(X_1, X_2) \wedge r(X_1) \wedge \neg g(X_1) \wedge \neg b(X_1) \wedge$$
$$\wedge \neg r(X_2) \wedge g(X_2) \wedge \neg b(X_2)] = \frac{1}{3}$$
$$P[\exists X_1, X_2 : F(X_1, X_2) \wedge r(X_1) \wedge \neg g(X_1) \wedge \neg b(X_1) \wedge$$
$$\wedge \neg r(X_2) \wedge \neg g(X_2) \wedge b(X_2)] = \frac{1}{3}$$
$$P[\exists X_1, X_2 : F(X_1, X_2) \wedge \neg r(X_1) \wedge g(X_1) \wedge \neg b(X_1) \wedge$$
$$\wedge \neg r(X_2) \wedge \neg g(X_2) \wedge b(X_2)] = \frac{1}{3}$$

(which can also be rewritten as probabilities of universally quantified formulas by negating the existentially quantified conjunctions). The global example Υ can be imagined as a complete directed graph (without self-loops) on 3 vertices v_1, v_2, v_3 where each of the vertices is colored by one of the "colors" r, g, b.

We claim that no distribution on global examples satisfies the above marginal probabilities for domain size greater than 3. This can be shown as follows. Using the intuitive view of Υ as a colored directed graph, the distribution on global examples of domain size e.g. 4 would be a distribution on graphs with 4 vertices. Such graphs would have to contain either two vertices not connected by an edge or two vertices connected by an edge but labeled with the same color or a vertex with no color. However, two such vertices would correspond to a local example which would otherwise have zero probability under the marginals estimated from Υ, which are shown above. While the above reasoning is for Model A, a similar argument can be used to show the same issues for Model B.

One of the consequences of the above example is that the unbiased estimates of relational marginals from Proposition 2 cannot always be used for defining distributions of arbitrary domain sizes. The section *Estimation* shows under which such unbiased estimates do exist, using the concept of *relational marginal polytopes*.

In order to construct distributions for arbitrary domain sizes, which have relational marginals that are close to the relational marginals given by some global example Υ, we will rely on the following construction which we call *expansion of global example*.
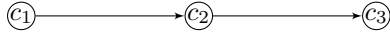
Definition 5 (Expansion of a global example). *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example where $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, and let l be a positive integer. Then the l-level expansion of Υ is a global example $\Upsilon' = (\mathcal{A}', \mathcal{C}')$ given by: $\mathcal{A}' = \{a\theta : a \in \mathcal{A}, \theta \in \Theta\}$, $\mathcal{C}' = \{c_1, c_2, \ldots, c_n, c_{n+1}, \ldots, c_{l \cdot n}\}$. Here, constants c_i, c_j are said to be congruent if $i \equiv j \mod n$. Here, $c_{n+1}, \ldots, c_{l \cdot n}$ are some arbitrary new constants and Θ is a set of all substitutions θ which satisfy the requirement that $c\theta$ is congruent with c for each $c \in \mathcal{C}$.*

Next we illustrate the notion of expansion of examples.

Example 6. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be given by*

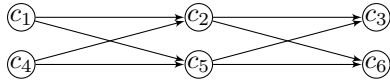$$\mathcal{A} = \{e(c_1, c_2), e(c_2, c_3)\}, \mathcal{C} = \{c_1, c_2, c_3\}.$$

Interpreting the predicate e as edge, the corresponding graph looks like:

Then the 2-level expansion of Υ is $\Upsilon' = (\mathcal{A}', \mathcal{C}')$ where

$$\mathcal{A} = \{e(c_1, c_2), e(c_2, c_3), e(c_4, c_5), e(c_5, c_6), e(c_1, c_5),$$
$$e(c_2, c_6), e(c_4, c_2), e(c_5, c_3)\}, \mathcal{C} = \{c_1, c_2, c_3, c_4, c_5, c_6\}.$$

Interpreting the predicate e again as edge, the expansion corresponds to the following graph:

The width-2 marginal probabilities on Υ and Υ' are:

$$P_{\Upsilon,2}(\omega_1) = \frac{1}{3} \quad P_{\Upsilon',2}(\omega_1) = \frac{7}{15} \qquad \omega_1 = (\{\}, \{1, 2\})$$

$$P_{\Upsilon,2}(\omega_2) = \frac{1}{3} \quad P_{\Upsilon',2}(\omega_2) = \frac{4}{15} \quad \omega_2 = (\{e(1,2)\}, \{1, 2\})$$

$$P_{\Upsilon,2}(\omega_3) = \frac{1}{3} \quad P_{\Upsilon',2}(\omega_3) = \frac{4}{15} \quad \omega_3 = (\{e(2,1)\}, \{1, 2\})$$

$$P_{\Upsilon,2}(\omega) = 0 \qquad P_{\Upsilon',2}(\omega) = 0 \qquad \omega \notin \{\omega_1, \omega_2, \omega_3\}$$

The differences between the marginal probabilities given by Υ and Υ' are at most $\frac{2}{15}$ in this case, which is quite high. However, it follows from what we show in turn that this is mostly because of the small size of Υ. For larger global examples, the difference between the marginals obtained from them and from their expansions will tend to be smaller.

Importantly, it is possible to bound the difference of the parameters obtained on expansions of global examples and the unbiased estimates obtained on the original examples.

Proposition 7. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example and Υ' its l-level expansion and let $n = |\mathcal{A}|$ and k be the width of local examples. Then for any formula α:*

$$|P_{\Upsilon,k}(\alpha) - P_{\Upsilon',k}(\alpha)| \;\leq\; 1 - \left(\frac{n - k + 1}{n}\right)^{k-1}$$

for Model A, and

$$|Q_{\Upsilon}(\alpha) - Q_{\Upsilon'}(\alpha)| \;\leq\; 1 - \left(\frac{n - k_\alpha + 1}{n}\right)^{k_\alpha - 1}$$

for Model B, where $k_\alpha = |vars(\alpha)|$.

Note that the difference between the true and modeled probabilities of a fixed formula α decays as $O\left(\frac{1}{n}\right)$.

The techniques described in this section have the following limitation. If we have a set of hard rules Φ_0 which are satisfied by a given Υ, these rules may not be satisfied in an expansion of Υ. This is not just a limitation of our method though. There are cases where it is not possible to extend a given Υ while satisfying the constraints (this is because we allow the use of equality in the formulas and because we use the unique names assumption). However, if the example Υ is large enough and it satisfies the hard rules, then the number of violations of these rules will be small, which follows again from Proposition 7.

In fact, it seems to be a desirable property that formulas α satisfying $P_{\Upsilon,k}(\alpha) = 1$ do not have to be treated as completely hard rules but as rules that "mostly" hold if they are learned from Υ, since it may be that they are not really rules that should always hold. Yet, if we actually took them as hard rules we would be forced to assign probability 0 to any example that violates them. It is possible to use the idea of expansions to obtain a distribution in which any formula α has nonzero probability and all properties of expansions are still preserved (such as those from Proposition 7). This can be achieved by randomly sampling additional atoms containing only congruent constants and adding them to the respective expansion. If we use a sufficiently high level of the expansion (at least k for Model A and at least $\max_{\alpha \in \Phi} |vars(\alpha)|$ for Model B) then the probability of any formula will be nonzero and not equal to one w.r.t. the distribution induced by the expansions with the sampled atoms.

Relational Marginal Polytopes

In this section, we define another important concept called *relational marginal polytope*, which will be used in the next section where we deal with estimation errors.

Definition 6 (Relational marginal polytope for Model A). *Let $k, m \in \mathbb{N}$ and $\Phi = \{\alpha_1, \ldots, \alpha_l\}$ be a set of formulas and Φ_0 be a set of hard rules. Let $\mathcal{C} = \{1, \ldots, m\}$ and $\mathcal{C}_k$ be the set of size-k subsets of $\mathcal{C}$. Then, for Model A, the relational marginal polytope of Φ of width k and cardinality m w.r.t. the hard rules from Φ_0 is the convex hull of the set*

$$\left\{ \left(\frac{\#_k(\alpha_1, \Upsilon)}{|\mathcal{C}_k|}, \ldots, \frac{\#_k(\alpha_l, \Upsilon)}{|\mathcal{C}_k|}\right) \Bigg| \Upsilon \in \Pi(\mathcal{C}, \Phi_0) \right\}.$$

Let Θ_{α_i} be the set of all injective substitutions from variables of α_i to constants from $\mathcal{C}$. Then, for Model B, the relational marginal polytope of Φ of cardinality m w.r.t. the hard rules from Φ_0 is the convex hull of the set

$$\left\{ \left(\frac{n(\alpha_1, \Upsilon)}{|\Theta_{\alpha_1}|}, \ldots, \frac{n(\alpha_l, \Upsilon)}{|\Theta_{\alpha_l}|}\right) \Bigg| \Upsilon \in \Pi(\mathcal{C}, \Phi_0) \right\}.$$

Any realizable set of relational marginals for Model A and Model B naturally corresponds to a point in the respective polytope. In the remainder of this paper, we only consider the cases when the relational marginal polytope is full-dimensional, that is, it does not live in a lower dimensional subspace which could happen if some of the relational marginals that define it were linearly dependent.

We will also need the concept of η-interior of a relational marginal. We say that a point y is in the η-interior of a relational marginal polytope P if P contains a ball of radius η with center in y. Using Proposition 4 and Proposition 7, we can show the following both for Model A and Model B.

Proposition 8. *Let θ be a vector representing the values of a set of relational marginals given by formulas from a set $\Phi = \{\alpha_1, \ldots, \alpha_l\}$. Let k be the width of the relational marginals of Model A or $k = \max_{\alpha_i} |vars(\alpha_i)|$ for Model B. Let the set of hard rules Φ_0 be empty. If θ is in the $\left(\eta + \sqrt{l}\left(1 - \left(\frac{m-k+1}{m}\right)^{k-1}\right)\right)$-interior of the relational marginal polytope of Φ of domain-size m then it is also in the η-interior of the relational marginal polytope of Φ for any domain size m'.*

Estimation

In this section, we present error bounds for the estimation of relational marginals. We start by defining the learning setting. Clearly, we need some assumptions on the training and test data and their relationship (otherwise one could always come up with a setting in which the error can be arbitrarily large). In order to stay close to realistic settings we assume that there is some large global example $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ that is not available and that represents the ground truth. That is what we essentially want to estimate for a given formula α is $P_{\aleph,k}(\alpha)$, but we do not have access to whole $\aleph$. Imagine for instance that $\aleph$ is the human gene regulatory network or Facebook. We assume that there is a process that samples size-m subsets of $\mathcal{C}_\aleph$ uniformly and that we have access to one such sample $\mathcal{C}_\Upsilon$ and also to the respective induced $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. So, for a given formula α, we need to estimate $P_{\aleph,k}(\alpha)$ using the available example Υ and the estimate needs to be realizable (otherwise the optimization problem would have no solution and the duality would also not hold). This is a reasonably realistic setting[5] as in practice we also do not have a distribution over different Facebooks but there is one Facebook and we want to model it based on a sample that is available to us.

We now provide theoretical upper bounds for the expected error of the estimates of $P_{\aleph,k}(\alpha)$ assuming the just described learning setting. However, we first need to describe the estimators. Based on the results from the previous sections, the estimator works as follows. Given a global example $\Upsilon = (\mathcal{A}_\Upsilon, \mathcal{C}_\Upsilon)$ and an integer n, which is the size of the domain of the modelled distribution (e.g. n can be size of $\aleph$'s domain if it is known), we construct the l-level expansion $\Upsilon^{(l)}$ of Υ, where $l = \lceil n/|\mathcal{C}_\Upsilon| \rceil$, and we use it to estimate the parameters as $\widehat{P}_\alpha = P_{\Upsilon^{(l)},k}(\alpha)$ for Model A and as $\widehat{Q}_\alpha = Q_{\Upsilon^{(l)}}(\alpha)$ for Model B. The following proposition introduces an upper bound for the expected error of the estimated parameters.

Proposition 9. *Let m and n be positive integers, α a closed formula and let k be the width of local examples. Let*

$\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ *be a global example, $\mathcal{C}_\Upsilon$ be sampled uniformly among all size-m subsets of $\mathcal{C}_\aleph$ and $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let $\widehat{A}_\aleph = P_{\aleph,k}(\alpha)$. Let $\widehat{B}_\Upsilon$ be an estimate computed from the l-level expansion of $\mathcal{C}_\Upsilon$. Then*

$$\mathbb{E}\left[\left|\widehat{A}_\aleph - \widehat{B}_\Upsilon\right|\right] \leq 1 - \left(\frac{m-k+1}{m}\right)^{k-1} + \sqrt{\frac{1 + 2\log 2}{4\lfloor m/k \rfloor}}.$$

In the case of model B, the same upper bound holds if we choose $k = |vars(\alpha)|$ and $\widehat{A}_\aleph = Q_\aleph(\alpha)$.

The proof of this proposition, which is contained in the online appendix, is based on a deviation bound for a randomized estimator of $\widehat{B}_\Upsilon$.

It is possible to improve the estimation in some cases. If the vector corresponding to the marginals Φ estimated from Υ is guaranteed to be in the $\left(\eta + \sqrt{l}\left(1 - \left(\frac{m-k+1}{m}\right)^{k-1}\right)\right)$-interior of the relational marginal polytope of domain of size $m = |\mathcal{C}_\Upsilon|$ for some $\eta > 0$, where k is the width of local examples for Model A (or $\max_{\alpha \in \Phi} |vars(\alpha)|$ for Model B), then, by Proposition 8, we can estimate the parameters directly from Υ without constructing its expansion. We then have the following improved bound:

$$\mathbb{E}\left[\left|\widehat{A}_\aleph - \widehat{B}_\Upsilon\right|\right] \leq \sqrt{(1 + 2\log 2)/(4\lfloor m/k \rfloor)}.$$

It is interesting to note that the lower-bound on effective sample size obtained from these bounds is $\lfloor m/k \rfloor$, which is also the maximum number of non-overlapping size-k subsets of $\mathcal{C}_\Upsilon$. A consequence for learning the parameters of models such as MLNs (which corresponds to relational marginal problems in Model B) is that this bound is inversely proportional to the number of variables in the used formulas, which also suggests an explanation for why learning with longer formulas is difficult.

Related Work

The relational marginals from Model A were recently introduced (Kuželka, Davis, and Schockaert 2017). However, they were only studied in a possibilistic setting, which differs substantially from the probabilistic maximum-entropy setting that we considered. The idea of using random substitutions (Model B) goes back to (Bacchus et al. 1992) who, however, only considered unary predicates. Schulte et al. (2014) used the random substitutions semantics to define a relational Bayesian network model for population statistics. However, their model is, not based on any underlying ground model, and it is unclear whether the distributions are always realizable by a ground model.

In the more restricted setting of exponential random graph models (ERGMs, Chatterjee and Diakonis 2013), a formally similar duality, based on densities of graph homomorphisms, has previously been established. To the best of our knowledge, however, such a duality has never been established in an SRL setting. In fact, even for ERGMs this duality has not yet been exploited for estimating parameters for models of different domain sizes, which is one of the key contributions of our work.

Certain statistical properties of learning have been already studied for SRL models. Xiang and Neville (2011) studied consistency but postulated rather strong assumptions[6], as a result of which their results are not comparable with ours. Their approach also differs in that it only considers distributions of labels conditioned on the underlying graph structure. It is interesting that a statistical estimation problem equivalent to the estimation of parameters in Model A has also been studied in the literature. In Nandi and Sen (1963), the variance of an estimator, equivalent to the unbiased estimator for Model A, was given. However, we are not aware of any work showing a deviation bound in the same setting, which was needed to establish the bound on expected error in Proposition 9. Interestingly, the effective sample size m/k stemming from the work (Nandi and Sen 1963) for variance of the estimator is almost the same as the effective sample size $\lfloor m/k \rfloor$ stemming from our deviation and expected-error bounds. This actually suggests that our bounds are rather tight. There were many works on U-statistics (Hoeffding 1948) which are related as well but they rely on assumptions that are generally not realistic for SRL and, in particular, are not applicable to our setting; the work of Nandi and Sen (1963) is an exception.

Conclusions

In this paper, we have introduced and studied relational marginal problems. Interestingly, this perspective enables learning a model that is applicable to data sets whose domain sizes differ from that of the training data. We established a relational counterpart of the classical duality between maximum-likelihood and max-entropy marginal problems. Then, we showed how to estimate and adjust parameters of the marginals in order to guarantee their realizability. We complemented these results by providing bounds on the expected errors of the estimates in a reasonable setting.

We believe that due to the simplicity and transparency of the learning setting that we introduced, this setting could play a similar role for SRL as the standard i.i.d. statistical learning setting plays for learning in propositional domains (Vapnik 1995). That is, as an idealized setting that is suitable for theoretical study, but that is not too far from settings that one encounters in reality. Still, it would be possible to extend the learning setting to make it more realistic. In particular, the sampling process that creates the training examples could be replaced by another sampling process that would take into account the structure of the relational data. That would probably make estimation of parameters and derivation of error bounds significantly more complex, and hence arguably less illuminating, which is why we leave it for future work.

Acknowledgment

Appendix: Duality

In this section we prove the duality results from the main text. First we recall the setting and notations. Let $\mathcal{C}$ be a set of constants, $\mathcal{A}$ the set of all atoms over $\mathcal{C}$ based on some given first-order language and let $\mathcal{C}_k$ denote the set of all k-element subsets of $\mathcal{C}$. Next, let Φ be a set of relational marginals and Φ_0 be a set of hard constraints, i.e. formulas α such that if $\Upsilon \not\models \alpha$ then $P(\Upsilon) = 0$. Let us also assume that there exists at least one distribution P' which is a model of Φ and which satisfies the positivity condition: $P'(\Upsilon) > 0$ for all Υ satisfying the hard constraints.

The optimization problem is then given by:

$$\sup_{\{P_\Upsilon : \Upsilon \in \Pi(\mathcal{C}, \Phi_0)\}} \sum_{\Upsilon \in \Pi_n(\mathcal{C}, \Phi_0)} P_\Upsilon \log \frac{1}{P_\Upsilon} \quad s.t. \quad (12)$$

$$\forall (\alpha_i, \theta_i) \in \Phi :$$
$$\frac{1}{|\mathcal{C}_k|} \sum_{\mathcal{S} \in \mathcal{C}_k} \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} \mathbb{1}(\Upsilon\langle\mathcal{S}\rangle \models \alpha_i) \cdot P_\Upsilon = \theta_i \quad (13)$$

$$\forall \Upsilon \in \Pi_n(\mathcal{C}, \Phi_0) : P_\Upsilon \geq 0, \quad \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} P_\Upsilon = 1 \quad (14)$$

Changing maximization of entropy to minimization of negative entropy, we obtain a convex optimization problem. Next, we construct the Lagrangian $L(P, \lambda, z)$ (here, P denotes the vector of all $P_\Upsilon \in \Pi(\mathcal{C}, \Phi_0)$, λ the vector of all λ_i's, and we will ignore the non-negativity constraints for the moment as it will turn out that they are enforced implicitly):

$$L(P, \lambda, z) = \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} P_\Upsilon \cdot \log P_\Upsilon -$$
$$- \frac{1}{|\mathcal{C}_k|} \sum_{\alpha_i \in \Phi} \sum_{\mathcal{S} \in \mathcal{C}_k} \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} \lambda_i \cdot \mathbb{1}(\Upsilon\langle\mathcal{S}\rangle \models \alpha_i) \cdot P_\Upsilon +$$
$$+ \sum_{\alpha_i \in \Phi} \lambda_i \cdot \theta_i + z - z \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} P_\Upsilon \quad (15)$$

To find the stationary points of $L(P, \lambda, z)$ w.r.t. P, we take the partial derivatives of the Lagrangian w.r.t. P_{Υ_j} and set them equal to zero:

$$\log P_{\Upsilon_j} - \frac{1}{|\mathcal{C}_k|} \sum_{\alpha_i \in \Phi} \sum_{\mathcal{S} \in \mathcal{C}_k} \lambda_i \cdot \mathbb{1}(\Upsilon_j\langle\mathcal{S}\rangle \models \alpha_i) - z - 1 = 0$$
$$(16)$$

Next we define $\#_k(\alpha, \Upsilon) = |\{\mathcal{S} \in \mathcal{C}_k : \Upsilon\langle\mathcal{S}\rangle \models \alpha\}|$. That is $\#_k(\alpha, \Upsilon)$ is a function counting the number of sets $\mathcal{S} \in \mathcal{C}_k$ such that the formula α is true in the restriction of

---

[6]The assumptions used in their work were *weak dependency* and bounded degree of graph nodes.

Υ to constants in $\mathcal{S}$ (i.e. those for which $\Upsilon\langle\mathcal{S}\rangle \models \alpha$). This allows us to rewrite (16) as follows:

$$\log P_{\Upsilon_j} - \sum_{\alpha_i \in \Phi} \lambda_i \frac{\#_k(\alpha_i, \Upsilon_j)}{|\mathcal{C}_k|} - z - 1 = 0 \quad (17)$$

From this, we get:

$$P_{\Upsilon_j} = \exp\left(z + 1 + \sum_{\alpha_i \in \Phi} \lambda_i \frac{\#_k(\alpha_i, \Upsilon_j)}{|\mathcal{C}_k|}\right) \quad (18)$$

Since we have $\sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} P_\Upsilon = 1$, $\exp(z+1)$ is a normalization constant.

By further algebraic manipulations, combining (15) and (18) we obtain the Lagrangian dual problem (where $\lambda_i \in \mathbb{R}$) is to maximize:

$$\sum_{\alpha_i \in \Phi} \lambda_i \theta_i - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} \lambda_i \frac{\#_k(\alpha_i, \Upsilon_j)}{|\mathcal{C}_k|}} \quad (19)$$

Due to the positivity assumption, Slater's condition (Boyd and Vandenberghe 2004) is satisfied and strong duality holds.

Remark 10. *The same reasoning can be also applied when we have multiple sets of relational marginals Φ_1, Φ_2, ..., Φ_n ($\Phi = \Phi_1 \cup \cdots \cup \Phi_n$) containing relational marginals of widths 1, 2, ..., n, respectively. The only difference is that the denominators $|\mathcal{C}_k|$ will vary according to widths (k_i) of the respective marginals (α_i) and we will denote the respective sets of subsets of $\mathcal{C}$ by $\mathcal{C}_{k_i}$.*

Let us perform a change of variables $w_i := \lambda_i/|\mathcal{C}_{k_i}|$ and denote the normalization constant of the distribution as Z. This gives us

$$P(\Upsilon) = \frac{1}{Z} \exp\left(\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon)\right) \quad (20)$$

for the distribution and

$$\sum_{\alpha_i \in \Phi} w_i \theta_i |\mathcal{C}_{k_i}| - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon)} \quad (21)$$

for the optimization criterion of the dual problem. Assuming that the marginals were estimated from a global example $\widehat{\Upsilon} \in \Pi(\mathcal{C}, \Phi_0)$ (note that here the domain $\mathcal{C}$ is the same as the domain of the global examples over which the distribution is defined) and that they still satisfy the positivity assumption, we can also rewrite (21) as

$$\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \widehat{\Upsilon}) - \log \sum_{\Upsilon \in \Pi(\mathcal{C}, \Phi_0)} e^{\sum_{\alpha_i \in \Phi} w_i \cdot \#_k(\alpha_i, \Upsilon)}$$

$$(22)$$

The reasoning is completely analogical for Model B. The only difference is that $\#_k(\alpha, \Upsilon)$ is replaced by $n(\alpha, \Upsilon)$ which counts the number of true groundings of α in Υ.

Appendix: Proofs

In this section we give proofs of the propositions from the main text.

Proposition 2. *Let $P(\Upsilon)$ be a distribution on domain size n and $k \leq n$ be an integer. Let $\Omega_\alpha = \{\Upsilon\langle\mathcal{S}\rangle : \mathcal{S} \subseteq \mathcal{C}, |\mathcal{S}| = k, \Upsilon\langle\mathcal{S}\rangle \models \alpha\}$ where $\Upsilon = (\mathcal{A}, \mathcal{C})$ is sampled according to the distribution $P(\Upsilon)$. Then, for a closed formula α,*

$$\widehat{p}_\alpha = |\Omega_\alpha| \cdot \binom{n}{k}^{-1}$$

is an unbiased estimate of $P_k(\alpha)$.

Proof. We have:

$$\mathbf{E}[\widehat{p}_\alpha] = \sum_{\Upsilon = (\mathcal{A}, \mathcal{C}) \in \Omega} P(\Upsilon) \cdot \binom{n}{k}^{-1} \cdot$$

$$\cdot |\{\mathcal{S} : \mathcal{S} \subseteq \mathcal{C}, |\mathcal{S}| = k, \Upsilon\langle\mathcal{S}\rangle\}| = \sum_{\Upsilon \in \Omega} P(\Upsilon) \cdot P_{\Upsilon, k}(\alpha) =$$

$$= \sum_{\omega : \omega \models \alpha} \sum_{\Upsilon \in \Omega} P(\Upsilon) \cdot P_{\Upsilon, k}(\omega) = P(\alpha).$$

$\square$

Proposition 4. *For Model A, if there is a width-k model of Φ on domain of size n then there is also a width-k model of Φ on domain of size m for any m satisfying $n \geq m \geq k$. For Model B, if there is a model of Φ on domain of size n then there is also a model of Φ on domain of size m for any m satisfying $n \geq m \geq \max_{\alpha \in \Phi} |vars(\alpha)|$.*

Proof. Let $P(.)$ be a model of Φ on domain of size n. Let $P'(.)$ be a distribution given by the following process: (i) sample $\Upsilon = (\mathcal{A}, \mathcal{C})$ according to $P(.)$, (ii) sample a subset $\mathcal{C}'$ from $\mathcal{C}$ uniformly among all subsets of $\mathcal{C}$ of cardinality m, and (iii) set $\Gamma := \Upsilon\langle\mathcal{C}'\rangle$. We claim that $P'(.)$ is also a width-k model of $\mathcal{M}$. To show that this is true it is enough to demonstrate that $P_k(\omega) = P'_k(\omega)$ for all width-k local examples ω, where $P_k(.)$ and $P'_k(.)$ are relational marginal distributions induced by $P(.)$ and $P'(.)$, respectively. This is straightforward to show. It suffices to notice that $P_k(\omega)$ does not change when we modify the relational marginal sampling process so that it first samples $\Upsilon = (\mathcal{A}, \mathcal{C})$, then it samples uniformly a subset $\mathcal{C}'$ of $\mathcal{C}$ of cardinality m, then it samples uniformly a subset $\mathcal{S}$ of $\mathcal{C}'$ and then a local example from $\Upsilon[\mathcal{S}]$. But this process is equivalent to sampling local examples from the global examples sampled according to the distribution $P'(.)$. Hence, $P_k(\omega) = P'_k(\omega)$.

The argument is completely analogical for Model B[7]. $\square$

It is worth noting here why the Proposition 4 would not hold for Model B if we did not require the sampled substitutions of Model B to be injective. In such a case, the distribution of local examples ω would not be the same if we first sampled a subset $\mathcal{C}'$ of size m from $\mathcal{C}$. For instance, for a formula α with two variables A and B, the proportion of

---

[7]Here, the condition $n \geq m \geq \max_{\alpha \in \Phi} |vars(\alpha)|$ is here so that we could actually sample injective substitutions.

samples mapping A and B to the same constant would be $\frac{1}{n}$ for sampling directly from $\mathcal{C}$ but $\frac{1}{m}$ when first sampling $\mathcal{C}'$ from $\mathcal{C}$ and then sampling the substitutions for $\mathcal{C}'$. So the two distributions would be no longer equivalent.

Lemma 1. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example and $\Upsilon' = (\mathcal{A}', \mathcal{C}')$ its l-level expansion. Let $n = |\mathcal{C}|$. Then there exists a distribution $\lambda(.)$ such that $P_{\Upsilon',k}(\omega) = (1-\gamma)P_{\Upsilon,k}(\omega) + \gamma\lambda(\omega)$. Here γ is the probability that a randomly sampled subset of k constants contains at least two congruent constants, i.e.*

$$\gamma = 1 - \frac{\binom{n}{k} \cdot l^k}{\binom{n \cdot l}{k}}$$

Proof. We prove this proposition for Model A; the arguments for Model B are essentially the same and therefore we omit them.

For notational convenience, let us define an indicator function $\mathbb{c}(\mathcal{S})$ which equals 1 when $\mathcal{S}$ contains two congruent constants and is 0 otherwise. We decompose the probability $P_{\Upsilon',k}(\omega)$ as:

$$P_{\Upsilon',k}(\omega) =$$
$$\frac{1}{|\Upsilon'[\mathcal{S}]|}P_{\mathcal{S}}[\omega \in \Upsilon'[\mathcal{S}]|\mathbb{c}(\mathcal{S}) = 1] \cdot P_{\mathcal{S}}[\mathbb{c}(\mathcal{S}) = 1]+$$
$$+ \frac{1}{|\Upsilon'[\mathcal{S}]|}P_{\mathcal{S}}[\omega \in \Upsilon'[\mathcal{S}]|\mathbb{c}(\mathcal{S}) = 0] \cdot P_{\mathcal{S}}[\mathbb{c}(\mathcal{S}) = 0]$$

where $P_{\mathcal{S}}[.]$ denotes probability under sampling of $\mathcal{S}$ according to Model A. Using elementary combinatorial reasoning, we have $P_{\mathcal{S}}[\mathbb{c}(\mathcal{S}) = 1] = \gamma$, $P_{\mathcal{S}}[\mathbb{c}(\mathcal{S}) = 0] = 1 - \gamma$ and we can set $\lambda(\omega) = \frac{1}{|\Upsilon'[\mathcal{S}]|}P_{\mathcal{S}}[\omega \in \Upsilon'[\mathcal{S}]|\mathbb{c}(\mathcal{S}) = 1]$. It remains to analyze $\frac{1}{|\Upsilon'[\mathcal{S}]|}P_{\mathcal{S}}[\omega \in \Upsilon'[\mathcal{S}]|\mathbb{c}(\mathcal{S}) = 0]$. First, we observe that this is the probability of the local example ω under the distribution "induced" by a uniform distribution over size-k subsets of $\mathcal{C}'$ which do not contain congruent constants. The same distribution can also be obtained as follows. First, pick uniformly at random a size-k subset $\mathcal{S}'$ of $\mathcal{C}$ and then assign, again uniformly and independently at random, to each of the constants in $\mathcal{S}'$ a constant sampled from the constants congruent to it in $\mathcal{C}'$. Then we may notice that it follows from the construction of expansions that the isomorphism class of the induced local example ω to be sampled is already determined when we sample the set $\mathcal{S}'$ and it is equal to $\Upsilon[\mathcal{S}']$. It follows that $\frac{1}{|\Upsilon'[\mathcal{S}]|}P_{\mathcal{S}}[\omega \in \Upsilon'[\mathcal{S}]|\mathbb{c}(\mathcal{S}) = 0] = P_{\Upsilon,k}(\omega)$ which finishes the proof. $\square$

Lemma 2. *Let P and λ be distributions over Ω and α be a closed and constant-free formula. Then*

$$\left| \sum_{\omega \in \Omega: \omega \models \alpha} P(\omega) - \lambda(\omega) \right| \leq 1$$

Proof. If $\sum_{\omega \in \Omega: \omega \models \alpha} P(\omega) - \lambda(\omega) \geq 0$ then we have

$$\sum_{\omega \in \Omega: \omega \models \alpha} P(\omega) - \lambda(\omega) \leq 1 - \sum_{\omega \in \Omega: \omega \models \alpha} \lambda(\omega) \leq 1$$

where the first and second inequality follow from the fact that $P(\omega)$ and $\lambda(\omega)$ most sum up to 1 over Ω and that they are positive. We can reason completely analogically for $\sum_{\omega \in \Omega: \omega \models \alpha} P(\omega) - \lambda(\omega) \leq 0$. The correctness of the lemma follows easily. $\square$

Proposition 7. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a global example and Υ' its l-level expansion and let $n = |\mathcal{A}|$ and k be the width of local examples. Then for any formula α:*

$$|P_{\Upsilon,k}(\alpha) - P_{\Upsilon',k}(\alpha)| \leq 1 - \left(\frac{n - k + 1}{n} \right)^{k-1}$$

for Model A, and

$$|Q_\Upsilon(\alpha) - Q_{\Upsilon'}(\alpha)| \leq 1 - \left(\frac{n - |vars(\alpha)| + 1}{n} \right)^{|vars(\alpha)| - 1}$$

for Model B.

Proof. We first prove the proposition for Model A. Let γ be as in Lemma 1. We have $|P_{\Upsilon,k}(\alpha) - P_{\Upsilon',k}(\alpha)| =$

$$\left| \sum_{\omega: \omega \models \alpha} (P_{\Upsilon,k}(\omega) - (1-\gamma) \cdot P_{\Upsilon,k}(\omega) - \gamma \cdot \lambda(\omega)) \right|$$

where $\lambda(.)$ is some suitable probability distribution. This can be rewritten as

$$|P_{\Upsilon,k}(\alpha) - P_{\Upsilon',k}(\alpha)| = \gamma \cdot \left| \sum_{\omega: \omega \models \alpha} (P_{\Upsilon,k}(\omega) - \lambda(\omega)) \right| \leq$$
$$\leq \gamma = 1 - \frac{l^k \cdot n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n \cdot l \cdot (n \cdot l - 1) \cdot \cdots \cdot (n \cdot l - k + 1)}$$

where the first inequality follows from Lemma 2. Next, we can verify that

$$1 - \frac{l^k \cdot n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n \cdot l \cdot (n \cdot l - 1) \cdot \cdots \cdot (n \cdot l - k + 1)} =$$
$$= 1 - \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n \cdot (n - \frac{1}{l}) \cdot \cdots \cdot (n - \frac{k+1}{l})}$$

is non-decreasing with l. Taking the limit $l \to \infty$ therefore preserves the inequality and we then get

$$|P_{\Upsilon,k}(\alpha) - P_{\Upsilon',k}(\alpha)| \leq 1 - \frac{n-1}{n} \cdot \frac{n-2}{n} \cdot \cdots \cdot \frac{n-k+1}{n}.$$

Now, to prove the proposition also for Model B, we proceed as follows. It is easy to verify that $Q_\Upsilon(\alpha)$ is equivalently given by the following process. First, we uniformly sample a local example ω of width $k = |vars(\alpha)|$ and then we uniformly sample a substitution ϑ from the set Θ_k of injective substitutions from $vars(\alpha)$ to the set $\{1, 2, \ldots, k\}$. $Q_\Upsilon(\alpha)$ is then also equal to the probability that $\omega \models \alpha\vartheta$. We can then exploit the arguments used to prove the result

for Model A. We have $|Q_\Upsilon(\alpha) - Q_{\Upsilon'}(\alpha)| =$

$$= \left| \frac{1}{|\Theta_k|} \sum_{\vartheta \in \Theta_k} \sum_{\omega : \omega \models \alpha\vartheta} P_{\Upsilon,k}(\omega) - P_{\Upsilon',k}(\omega) \right|$$

$$\leq 1 - \left(\frac{n-k+1}{n} \right)^{k-1}$$

$$= 1 - \left(\frac{n - |vars(\alpha)| + 1}{n} \right)^{|vars(\alpha)|-1}$$

where the inequality follows from the same algebraic manipulations that we performed above for Model A. $\square$

Proposition 8. *Let θ be a vector representing the values of a set of relational marginals given by formulas from a set $\Phi = \{\alpha_1, \dots, \alpha_l\}$. Let k be the width of the relational marginals of Model A or $k = \max_{\alpha_i} |vars(\alpha_i)|$ for Model B. Let the set of hard rules Φ_0 be empty. If θ is in the $\left(\eta + \sqrt{l}\left(1 - \left(\frac{m-k+1}{m}\right)^{k-1}\right)\right)$-interior of the relational marginal polytope of Φ of domain-size m then it is also in the η-interior of the relational marginal polytope of Φ for any domain size m'.*

Proof. Let us denote by $S(\Upsilon_i)$ the vector of values of the relational marginals corresponding to the formulas $\alpha_1, \dots, \alpha_l$ computed from Υ_i. Let B be a point on the boundary of the relational marginal polytope. It follows from the definitions that we can write B as a linear combination[8] $B = a_1 \cdot S(\Upsilon_1) + \cdots + a_l \cdot S(\Upsilon_l)$ where all a_i's are positive and $a_1 + \cdots + a_l = 1$ for some suitable $\Upsilon_1, \dots, \Upsilon_l$ on domain of size m. Let $\Upsilon'_1, \dots, \Upsilon'_l$ be the expansions of the respective $\Upsilon_1, \dots, \Upsilon_l$ (we pick the level of these expansions sufficient to make the cardinality of their domains greater than the desired cardinality m') and let $B' = a_1 \cdot S(\Upsilon'_1) + \cdots + a_l \cdot S(\Upsilon'_l)$. For notational convenience, let us denote

$$\xi = 1 - \left(\frac{n-k+1}{n} \right)^{k-1}.$$

With this notation we have that θ is in the $(\eta + \xi\sqrt{l})$-interior of the marginal polytope of domain of size m. Then, using Proposition 7, we have (for both Model A and Model B):

$$\|S(\Upsilon_i) - S(\Upsilon'_i)\|_2 \leq \xi\sqrt{l}$$

where we note that l is the number of formulas in Φ. Using triangle inequality, it follows that

$$\|B - B'\|_2 =$$
$$= \|a_1(S(\Upsilon_1) - S(\Upsilon'_1)) + \cdots + a_l(S(\Upsilon_l) - S(\Upsilon'_l))\|_2$$
$$\leq a_1\|S(\Upsilon_1) - S(\Upsilon'_1)\|_2 + \cdots + a_l\|S(\Upsilon_l) - S(\Upsilon'_l)\|_2$$
$$\leq (a_1 + \cdots + a_l) \cdot \xi \cdot \sqrt{l} = \xi\sqrt{l}$$

---

[8]In particular, this follows from the fact that relational marginal polytope is a convex hull of a finite set of points in l-dimensional space where the points are given by $S(\Upsilon)$ for $\Upsilon \in \Pi(\mathcal{C}, \emptyset)$.

Intuitively this means that any point on the boundary of the polytope for domain size m will have a point on the boundary or inside the polytope for domain size m' in distance not greater than $\xi\sqrt{l}$. Since the point θ is in the $(\eta + \xi\sqrt{l})$-interior of the polytope for domain size m, it follows that it must also be in the η-interior of the polytope for domain size m' (for any m'), which is what we needed to show. $\square$

Next we prove Proposition 9 using a series of lemmas. We first state and prove the lemmas for Model A and after that, separately, also their analogues for Model B.

Lemma 3 (Model A Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle\mathcal{C}_\Upsilon\rangle$. Let $\mathbf{X} = (\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{\lfloor \frac{n}{k} \rfloor})$ be a vector of subsets of $\mathcal{C}_\aleph$, each sampled uniformly and independently of the others from all size-k subsets of $\mathcal{C}_\aleph$. Next let $\mathbf{Y} = (\mathcal{S}'_1, \mathcal{S}'_2, \dots, \mathcal{S}'_{\lfloor \frac{n}{k} \rfloor})$ be a vector sampled by the following process:*

1. *Let $\mathcal{I}' = \{1, 2, \dots, |\mathcal{C}_\aleph|\}$.*
2. *Sample subsets $\mathcal{I}'_1, \dots, \mathcal{I}'_{\lfloor \frac{n}{k} \rfloor}$ of size k uniformly from $\mathcal{I}'$.*
3. *Sample an injective function $g : \bigcup_{\mathcal{I}'_i \subseteq \mathcal{I}'} \mathcal{I}'_i \to \mathcal{C}_\Upsilon$ uniformly from all such functions.*
4. *Let $\mathcal{S}'_i = g(\mathcal{I}'_i)$ for all $0 \leq i \leq \lfloor \frac{n}{k} \rfloor$.*

Then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Proof. (Sketch) Let h be a bijection from $\mathcal{I}$ to $\mathcal{C}_\aleph$ sampled uniformly from the set of all such bijections. If we sample $\mathcal{I}_1, \dots, \mathcal{I}_{\lfloor \frac{n}{k} \rfloor}$ of size k uniformly from $\mathcal{I}$ then the distribution of $(h(\mathcal{I}_1), \dots, h(\mathcal{I}_{\lfloor \frac{n}{k} \rfloor}))$ will be the same as that of $\mathbf{X}$. In fact, most of h is actually irrelevant and we can replace sampling h by uniformly sampling an injective function $g : \bigcup \mathcal{I}_i \to \mathcal{C}_\aleph$ after we sample the sets $\mathcal{I}_i$. Finally, it is easy to realize that the function g can also be sampled as follows. First, sample a subset $\mathcal{C}$ of $\mathcal{C}_\aleph$ of size n uniformly and then sample uniformly an injective function g such that $\text{Range}(g) \subseteq \mathcal{C}$. The reason why we can do this is because $|\bigcup \mathcal{I}_i| \leq n$. This finishes the proof of the lemma as this is actually the same as the process of sampling $\mathbf{Y}$. $\square$

Lemma 4 (Model A Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle\mathcal{C}_\Upsilon\rangle$. Let $\mathbf{Y}$ be sampled as in Lemma 3 (i.e. $\mathbf{Y}$ is sampled only using Υ and not directly $\aleph$). Let α be a closed and constant-free formula. Let*

$$\widehat{A}_\Upsilon = \frac{1}{\lfloor \frac{n}{k} \rfloor} \sum_{\mathcal{S}'_i \in \mathbf{Y}} \mathbb{1}(\Upsilon\langle\mathcal{S}'_i\rangle \models \alpha)$$

and let $A_\aleph = P_{\aleph,k}(\alpha)$. Then we have

$$P\left[\left| \widehat{A}_\Upsilon - A_\aleph \right| \geq \epsilon \right] \leq 2\exp\left(-2\left\lfloor \frac{n}{k} \right\rfloor \epsilon^2 \right).$$

Proof. We define an estimator on the set $\mathbf{X}$ which is sampled as in Lemma 3

$$\widehat{A} = \frac{1}{\left\lfloor \frac{n}{k} \right\rfloor} \sum_{\mathcal{S}_i \in \mathbf{X}} \mathbb{1}(\aleph\langle \mathcal{S}_i \rangle \models \alpha).$$

By Lemma 3, we know that

$$P\left[\left| \widehat{A} - A_\aleph \right| \geq \epsilon \right] = P\left[\left| \widehat{A}_\Upsilon - A_\aleph \right| \geq \epsilon \right],$$

so we only need to show

$$P\left[\left| \widehat{A} - A_\aleph \right| \geq \epsilon \right] \leq 2 \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor \epsilon^2 \right).$$

Since $\mathcal{S}_i$ in $\mathbf{X}$ are independent, the inequality above follows immediately from the fact that $\mathbb{E}\left[\widehat{A} \right] = A_\aleph$ (which follows $\mathbb{E}\left[\mathbb{1}(\aleph\langle \mathcal{S}_i \rangle \models \alpha) \right] = A_\aleph$ and the linearity property of expectation) and the Chernoff-Hoeffding theorem. $\square$

Lemma 5 (Model A Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let $\mathbf{Y}$ be sampled as in Lemma 3 (i.e. $\mathbf{Y}$ is sampled only using Υ and not directly $\aleph$). Let α be a closed and constant-free formula. Let*

$$\widehat{A}_\Upsilon = \frac{1}{\left\lfloor \frac{n}{k} \right\rfloor} \sum_{\mathcal{S}'_i \in \mathbf{Y}} \mathbb{1}(\Upsilon\langle \mathcal{S}'_i \rangle \models \alpha)$$

and let $A_\aleph = P_{\aleph,k}(\alpha)$. Then we have

$$\mathbb{E}\left[\left| \widehat{A}_\Upsilon - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.$$

Proof. Let $Z := \left| \widehat{A}_\Upsilon - A_\aleph \right|$. First, notice that

$$\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]},$$

so we now try to bound $\mathbb{E}\left[Z^2 \right]$. To bound $\mathbb{E}\left[Z^2 \right]$, note that for any $u > 0$,

$$\begin{aligned}
\mathbb{E}\left[Z^2 \right] &= \int_0^\infty P\left[Z^2 \geq t \right] dt \\
&= \int_0^\infty P\left[Z \geq \sqrt{t} \right] dt \\
&\leq u + \int_u^\infty P\left[Z \geq \sqrt{t} \right] dt \\
&\leq u + 2 \int_u^\infty \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor t \right) dt \\
&= u + \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor u \right) / \left\lfloor \frac{n}{k} \right\rfloor.
\end{aligned}$$

Since u was arbitrary, we may choose it to minimize the obtained bound. The optimal choice is $u = \frac{\log 2}{2\lfloor n/k \rfloor}$, which yields $\mathbb{E}\left[Z^2 \right] \leq \frac{1 + 2\log 2}{4\lfloor n/k \rfloor}$. $\square$

Lemma 6 (Model A Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let α be a closed and constant-free formula and let $\mathcal{C}_k$ denote all size-k subsets of $\mathcal{C}_\Upsilon$. Let*

$$\widetilde{A}_\Upsilon = \binom{n}{k}^{-1} \sum_{\mathcal{S} \in \mathcal{C}_k} \mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha)$$

and let $A_\aleph = P_{\aleph,k}(\alpha)$. Then we have

$$\mathbb{E}\left[\left| \widetilde{A}_\Upsilon - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.$$

Proof. First we define an auxiliary estimator $\widetilde{A}_\Upsilon^{(q)}$. Let $\mathbf{Y}_q$ be a vector of $n \cdot q$ size-k subsets of $\mathcal{C}_\Upsilon$ where the subsets of $\mathcal{C}_\Upsilon$ are sampled in the same way as the elements of the vector $\mathbf{Y}$ in Lemma 3. Let us define

$$\widetilde{A}_\Upsilon^{(q)} = \frac{1}{q \cdot n} \sum_{\mathcal{S} \in \mathbf{Y}_q} \mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha).$$

Using the triangle inequality and Lemma 5, we can show that

$$\mathbb{E}\left[\left| \widetilde{A}_\Upsilon^{(q)} - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.$$

We have (again using triangle inequality)

$$\begin{aligned}
\mathbb{E}\left[\left| \widetilde{A}_\Upsilon - A_\aleph \right| \right] &= \mathbb{E}\left[\left| \widetilde{A}_\Upsilon - \widetilde{A}_\Upsilon^{(q)} + \widetilde{A}_\Upsilon^{(q)} - A_\aleph \right| \right] \\
&\leq \mathbb{E}\left[\left| \widetilde{A}_\Upsilon - \widetilde{A}_\Upsilon^{(q)} \right| \right] + \mathbb{E}\left[\left| \widetilde{A}_\Upsilon^{(q)} - A_\aleph \right| \right] \\
&\leq \mathbb{E}\left[\left| \widetilde{A}_\Upsilon - \widetilde{A}_\Upsilon^{(q)} \right| \right] + \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.
\end{aligned}$$

It follows from the strong law of large numbers (which holds for any Υ) and Proposition 2 that $P[\lim_{q \to \infty} \widetilde{A}_\Upsilon^{(q)} = \widetilde{A}_\Upsilon] = 1$. Since q was arbitrary, we can use this to conclude

$$\mathbb{E}\left[\left| \widetilde{A}_\Upsilon - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.$$

$\square$

Next we state the analogues of the above lemmas also for Model B (for the first reading, we recommend skipping these lemmas and proceeding directly to the proof of Proposition 9). We omit the proofs as they are completely analogical to their respective counterparts for Model A (we just replace sampling of subsets by sampling of injective substitutions).

Lemma 7 (Model B Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$.*

Let $\mathbf{X} = (\vartheta_1, \vartheta_2, \ldots, \vartheta_{\lfloor \frac{n}{k} \rfloor})$ be a vector of injective substitutions from a given size-k set of variables $\mathcal{V}$ to $\mathcal{C}_\aleph$, each sampled uniformly and independently of the others. Next let $\mathbf{Y} = (\vartheta'_1, \vartheta'_2, \ldots, \vartheta'_{\lfloor \frac{n}{k} \rfloor})$ be a vector sampled by the following process:

1. Let $\mathcal{I}' = \{1, 2, \ldots, |\mathcal{C}_\aleph|\}$.
2. Sample **ordered** subsets $\mathcal{I}'_1, \ldots, \mathcal{I}'_{\lfloor \frac{n}{k} \rfloor}$ of size k uniformly from $\mathcal{I}'$.
3. Sample an injective function $g : \bigcup_{\mathcal{I}'_i \subseteq \mathcal{I}'} \mathcal{I}'_i \to \mathcal{C}_\Upsilon$ uniformly from all such functions.
4. Let $\vartheta'_i = \{V_j \mapsto c_j | V_j \in \mathcal{V} \text{ and } (c_1, \ldots, c_{|\mathcal{V}|}) = g(\mathcal{I}'_i)\}$ for all $0 \leq i \leq \lfloor \frac{n}{k} \rfloor$, i.e. ϑ'_i is an injective substitution from $\mathcal{V}$ to $\mathcal{C}_\Upsilon$.

Then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Lemma 8 (Model B Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let α be a closed and constant-free formula. Let $\mathbf{Y}$ be sampled as in Lemma 7 (i.e. $\mathbf{Y}$ is sampled only using Υ and not directly $\aleph$), where we set $\mathcal{V} = vars(\alpha)$. Let*

$$\widehat{A}_\Upsilon = \frac{1}{\lfloor \frac{n}{k} \rfloor} \sum_{\vartheta'_i \in \mathbf{Y}} \mathbb{1}(\Upsilon \models \alpha \vartheta'_i)$$

and let $A_\aleph = Q_\aleph(\alpha)$. Then we have

$$P\left[\left| \widehat{A}_\Upsilon - A_\aleph \right| \geq \epsilon \right] \leq 2 \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor \epsilon^2 \right).$$

Lemma 9 (Model B Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let $\mathbf{Y}$ be sampled as in Lemma 7 (i.e. $\mathbf{Y}$ is sampled only using Υ and not directly $\aleph$), where we set $\mathcal{V} = vars(\alpha)$. Let α be a closed and constant-free formula. Let*

$$\widehat{A}_\Upsilon = \frac{1}{\lfloor \frac{n}{k} \rfloor} \sum_{\vartheta'_i \in \mathbf{Y}} \mathbb{1}(\Upsilon \models \alpha \vartheta'_i)$$

and let $A_\aleph = Q_\aleph(\alpha)$. Then we have

$$\mathbb{E}\left[\left| \widehat{A}_\Upsilon - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/k \rfloor}}.$$

Lemma 10 (Model B Case). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$, where $\mathcal{C}_\aleph$ is a set of constants and $\mathcal{A}_\aleph$ is a set of ground atoms involving only constants from $\mathcal{C}_\aleph$ and let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let α be a closed and constant-free formula and Θ_α denote all injective substitutions from $vars(\alpha)$ to $\mathcal{C}_\Upsilon$. Let*

$$\widetilde{A}_\Upsilon = \frac{1}{|\Theta_\alpha|} \sum_{\vartheta \in \Theta_\alpha} \mathbb{1}(\Upsilon \models \alpha \vartheta)$$

and let $A_\aleph = Q_\aleph(\alpha)$. Then we have

$$\mathbb{E}\left[\left| \widetilde{A}_\Upsilon - A_\aleph \right| \right] \leq \sqrt{\frac{1 + 2\log 2}{4\lfloor n/|vars(\alpha)| \rfloor}}.$$

Now we have all the ingredients needed to prove Proposition 9.

Proposition 9 *Let m and n be positive integers, α a closed formula and let k be the width of local examples. Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a global example, $\mathcal{C}_\Upsilon$ be sampled uniformly among all size-m subsets of $\mathcal{C}_\aleph$ and $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let $\widehat{A}_\aleph = P_{\aleph,k}(\alpha)$. Let $\widehat{B}_\Upsilon$ be an estimate computed from the l-level expansion of $\mathcal{C}_\Upsilon$. Then*

$$\mathbb{E}\left[\left| \widehat{A}_\aleph - \widehat{B}_\Upsilon \right| \right] \leq 1 - \left(\frac{m - k + 1}{m} \right)^{k-1} + \sqrt{\frac{1 + 2\log 2}{4\lfloor m/k \rfloor}}.$$

In the case of model B, the same upper bound holds if we choose $k = |vars(\alpha)|$ and $\widehat{A}_\aleph = Q_\aleph(\alpha)$.

Proof. Let $\widehat{A}_\Upsilon = P_{\Upsilon,k}(\alpha)$ ($\widehat{A}_\Upsilon = Q_\Upsilon(\alpha)$, respectively). Then we have

$$\mathbb{E}\left[\left| \widehat{A}_\aleph - \widehat{B}_\Upsilon \right| \right] = \mathbb{E}\left[\left| \widehat{A}_\aleph - \widehat{A}_\Upsilon + \widehat{A}_\Upsilon - \widehat{B}_\Upsilon \right| \right]$$

$$\leq \mathbb{E}\left[\left| \widehat{A}_\Upsilon - \widehat{B}_\Upsilon \right| \right] + \mathbb{E}\left[\left| \widehat{A}_\aleph - \widehat{A}_\Upsilon \right| \right]$$

$$\leq 1 - \left(\frac{m - k + 1}{m} \right)^{k-1} + \sqrt{\frac{1 + 2\log 2}{4\lfloor m/k \rfloor}}$$

where the last inequality follows from Proposition 7 and Lemma 6 for Model A, and Lemma 10 for Model B, respectively. $\square$

References

Bacchus, F.; Grove, A. J.; Koller, D.; and Halpern, J. Y. 1992. From statistics to beliefs. In *Proceedings of the 10th National Conference on Artificial Intelligence*, 602–608.

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

Chatterjee, S., and Diaconis, P. 2013. Estimating and understanding exponential random graph models. *The Annals of Statistics* 41(5):2428–2461.

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

Hoeffding, W. 1948. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics* 293–325.

Jain, D.; Kirchlechner, B.; and Beetz, M. 2007. Extending Markov logic to model probability distributions in relational domains. In *KI 2007: Advances in Artificial Intelligence, 30th Annual German Conference on AI, KI 2007*, 129–143.

Kuželka, O.; Davis, J.; and Schockaert, S. 2017. Induction of interpretable possibilistic logic theories from relational data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, 1153–1159.

Nandi, H., and Sen, P. 1963. On the properties of u-statistics when the observations are not independent: Part two unbiased estimation of the parameters of a finite population. *Calcutta Statistical Association Bulletin* 12(4):124–148.

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

Schulte, O.; Khosravi, H.; Kirkpatrick, A. E.; Gao, T.; and Zhu, Y. 2014. Modelling relational statistics with Bayes nets. *Machine Learning* 94(1):105–125.

Shalizi, C. R., and Rinaldo, A. 2013. Consistency under sampling of exponential random graph models. *The Annals of Statistics* 41(2):508–535.

Singh, M., and Vishnoi, N. K. 2014. Entropy, optimization and counting. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 50–59. ACM.

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

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

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

Xiang, R., and Neville, J. 2011. Relational learning with one network: An asymptotic analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, 779–788.