PAC-Reasoning in Relational Domains

Ondřej Kuželka
Department of CS
KU Leuven
Leuven, Belgium

Yuyi Wang
Disco Group
ETH Zurich
Zurich, Switzerland

Jesse Davis
Department of CS
KU Leuven
Leuven, Belgium

Steven Schockaert
School of CS & Informatics
Cardiff University
Cardiff, UK

Abstract

We consider the problem of predicting plausible missing facts in relational data, given a set of imperfect logical rules. In particular, our aim is to provide bounds on the (expected) number of incorrect inferences that are made in this way. Since for classical inference it is in general impossible to bound this number in a non-trivial way, we consider two inference relations that weaken, but remain close in spirit to classical inference.

1 INTRODUCTION

In this paper we study several forms of logical inference for predicting plausible missing facts in relational data. While a variety of approaches have already been studied for this task, ranging from (relational versions of) probabilistic graphical models [19, 4] to neural-network architectures [24, 20] and graph-based methods [15, 16], logic-based inference has several advantages over these other forms of inference. For example, logic-based inference is explainable: there is a proof for any derived statement, which can, in principle, be shown to a human user. It is also more transparent than most other methods, in the sense that a knowledge base as a whole can be understood and modified by domain experts. On the other hand, classic logical inference can be very brittle when some of the rules which are used are imperfect, or some of the initial facts may be incorrect.

Statistical relational learning approaches, such as Markov logic networks [19] or probabilistic logic programming [4], offer a solution to this latter problem, but they require learning a joint probability distribution over the set of possible worlds. This distribution is typically estimated from one or several large examples using maximum likelihood, which essentially corresponds to finding a maximum-entropy distribution given by a set of sufficient statistics. However, there are usually no guarantees on the learned distributions beyond guarantees for the sufficient statistics (see, e.g., [12]), which means that we do not have much control over the quality of the predictions. Moreover, these models are not easy to modify, and are not always easy to explain because the way in which probabilities are computed can simply be too complex.

In this paper we focus on forms of inference that stay as close to classical logic as possible while not breaking completely when the given theory happens to be "mildly" inconsistent with the data. This problem of reasoning under inconsistency has a long tradition in the field of artificial intelligence, with common solutions including the use of paraconsistent logics [3, 18], belief revision [8] (and related inconsistency repair mechanisms [11]), and argumentation-based inference [7, 2]. In contrast to these approaches, however, our specific aim is to study forms of inference that can allow us to bound the (expected) number of mistakes that are made. To this end, we introduce two inference relations called *k-entailment* and *voting entailment*, both of which are close to classical logic, and in particular do not require rules to be weighted. We define them such that errors produced by imperfect rules would not necessarily propagate too much in the given relational data.

As our main contribution, we are able to show that in a relational learning scenario from [12], in which a (large) training example and a test example are sampled from a hidden relational structure, there are non-trivial PAC-type bounds on the number of errors that a theory learned on the training example produces on the test example. From this perspective, our work can also be seen as a relational-learning counterpart of PAC semantics [23].

Technical contributions. The results presented in this paper rest mainly on the following two technical contributions: (i) the introduction of bounds on the worst case behavior of the considered inference relations, and (ii) new concentration inequalities for sampling from relational data without replacement that allow us to bound the (expected) test error as a function of the training error, in the spirit of classical PAC-learning results [22].

2 PRELIMINARIES

In this paper we consider 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. 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. 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.

3 PROBLEM SETTING

First we describe the learning setting considered in this paper. It follows the setting from [12],which was used to study the estimation of relational marginals.

An 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 $\mathcal{C}$. An example is intended to provide a complete description of the world, hence any ground atom over $\mathcal{C}$ which is not contained in $\mathcal{A}$ is implicitly assumed to be false. Note that this is why we have to explicitly specify $\mathcal{C}$, as opposed to simply considering the set of constants appearing in $\mathcal{A}$.

In practice, we usually only have partial information about some example of interest. The problems we consider in this paper relate to how we can then reason about the probability that a given ground atom is true (i.e. belongs to the example). To estimate such probabilities, we assume that we are given a fragment of the example, which we can use as training data. Specifically, 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 $\mathcal{S}$. In a given example, any closed formula α is either true or false. To assign probabilities to formulas in a meaningful way, we consider how often the formula is satisfied in small fragments of the given example.

Definition 1 (Probability of a formula [12])**.** *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example and $k \in \mathbb{N}$. For a closed formula α without constants, we define its probability as follows*[1]*:*

$$Q_{\Upsilon,k}(\alpha) = P_{\mathcal{S} \sim Unif(\mathcal{C},k)} \left[\Upsilon\langle \mathcal{S} \rangle \models \alpha \right]$$

where $Unif(\mathcal{C}, k)$ denotes uniform distribution on size-k subsets of $\mathcal{C}$.

Clearly $Q_{\Upsilon,k}(\alpha) = \frac{1}{|\mathcal{C}_k|} \cdot \sum_{\mathcal{S} \in \mathcal{C}_k} \mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha)$ where $\mathcal{C}_k$ is the set of all size-k subsets of $\mathcal{C}$.

The above definition is also extended straightforwardly to probabilities of sets of formulas (which we will also call *theories* interchangeably). If Φ is a set of formulas, we set $Q_{\Upsilon,k}(\Phi) = Q_{\Upsilon,k}(\bigwedge \Phi)$ where $\bigwedge \Phi$ denotes the conjunction of all formulas in Φ.

Example 1. *Let $sm/1$ be a unary predicate denoting that someone is a smoker, e.g. $sm(alice)$ means that alice is a smoker. Let us have an example $\Upsilon = (\{fr(alice, bob), sm(alice), sm(eve)\}, \{alice, bob, eve\})$, and formulas $\alpha = \forall X : sm(X)$ and $\beta = \exists X, Y : fr(X, Y)$. Then, for instance, $Q_{\Upsilon,1}(\alpha) = 2/3$, $Q_{\Upsilon,2}(\alpha) = 1/3$ and $Q_{\Upsilon,2}(\beta) = 1/3$.*

Definition 2 (Masking)**.** *A masking process is a function κ from examples to ground conjunctions that assigns to any $\Upsilon = (\mathcal{A}, \mathcal{C})$ a conjunction of ground literals β such that $\Upsilon \models \beta$. We also define $\kappa(\Upsilon)\langle \mathcal{S} \rangle$ to be the conjunction consisting of all literals from $\kappa(\Upsilon)$ that contain only constants from $\mathcal{S}$.*

Unlike examples, masked examples only encode partial information about the world. This is why they are encoded using conjunctions of literals, so we can explicitly encode which atoms we know to be false.

Example 2. *Let $\Upsilon = \{sm(alice), fr(alice, bob), \{alice, bob\}\}$. Then a masking process κ may, for instance, yield $\kappa(\Upsilon) = \neg sm(bob) \wedge sm(alice)$. In this case $\kappa(\Upsilon)$ retains the information that alice is a smoker and bob is not, but it no longer contains any information about their friendship relation.*

Next we introduce the statistical setting considered in this paper.

Definition 3 (Learning setting)**.** *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be an example and κ be a masking function. Let $\mathcal{C}_\Upsilon \subseteq \mathcal{C}_\aleph$ and $\mathcal{C}_\Gamma \subseteq \mathcal{C}_\aleph$ be uniformly sampled subsets of size n and u, respectively. We call $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$ the* training example *and $\Gamma = \aleph\langle \mathcal{C}_\Gamma \rangle$ the* test example. *We assume that the learner receives Υ in the training phase and $\kappa(\Gamma)$ in the test phase.*

With slight abuse of terminology, we will sometimes say that Υ and Γ are sampled from $\aleph$.

[1] We will use Q for probabilities of formulas as defined in this section, to avoid confusion with other "probabilities" we deal with in the text.

In addition to the training example Υ and masked test example $\kappa(\Gamma)$, we will assume that we are given a set of formulas Φ (which we will also refer to as rules). Our main focus will be on how these formulas can be used to recover as much of Γ as possible. Rather than specifying a loss function that should be minimized, we want to find a form of inference which allows us to provide bounds on the (expected) number of incorrect literals that can be inferred from $\{\kappa(\Gamma)\} \cup \Phi$. Note that in this case, the training example Υ is used to estimate the accuracy of the set of formulas. We also analyze the case where the rules are learned from the training example Υ (in the spirit of classical PAC-learning results).

Among others, the setting from Definition 3 is close to how Markov logic networks are typically used. For instance, when training Markov logic networks, one typically starts with a training example that contains all facts (i.e. nothing is unknown about the training set), on which a model is trained. This model is then used to predict unknown facts about a test example. However, unlike for Markov logic networks, we do not attempt to learn a probability distribution. It was shown in [14] that models based on classic logical inference, like those considered in this paper, work well in practice for relational inference from evidence sets containing a *small* number of constants (domain elements). Thus, such models are also of considerable practical interest.

4 REASONING WITH INACCURATE RULES

When reasoning with imperfect rules, using classical inference can have drastic consequences, as we will illustrate in Section 4.1. Even a single mistake can lead to many errors, since an incorrectly derived literal can be used as the basis for further inferences. This means that classical inference is not suitable for the considered setting, even in cases where the given rules have perfect accuracy on the training example. Intuitively, to allow for any meaningful bounds to be derived, we need to prevent arbitrarily long chains of inference. To this end, we propose and motivate the use of a restricted form of inference, called k-entailment, in Section 4.2. A further restriction on inferences, based on a form of voting, is subsequently discussed in Section 4.3. In Section 5 we will then show which bounds can be derived for these two restricted forms of inference.

4.1 WHEN CLASSICAL REASONING LEADS TO ERRORS

The next example, which is related to label propagation as studied e.g. in [26], shows that classic logical reasoning on the obtained relational sample may produce *many* mistakes even when all the available rules are very *accurate*.

Example 3. *Let $k = 2$, $\Gamma = \{\{rare(c_1)\}, \{c_1, c_2, \dots, c_{1000000}\}$, and $\alpha = \forall X, Y : rare(X) \Rightarrow rare(Y)$. While the rule does not intuitively make sense, its accuracy is actually very high $Q_{\Gamma,k}(\alpha) = 1 - 999999/(0.5 \cdot 1000000 \cdot 999999) = 0.999998$. When we apply this rule with the evidence $rare(c_1)$, we derive $rare(c_2), \dots, rare(c_{1000000})$, all of which are incorrect (i.e. not included in Γ).*

Note that in this paper, we are interested in worst-case behavior, in the sense that the masking process which is used may be seen as adversarial. The next example further illustrates how adversarial masking processes can lead to problems, even for rules with near-perfect accuracy.

Example 4. *Let $k = 2$, $\Gamma = \{\{rare(c_1), e(c_1, c_2), e(c_2, c_3), \dots, e(c_{999999}, c_{1000000})\}, \{c_1, c_2, \dots, c_{1000000}\}$, and $\alpha = \forall X, Y : rare(X) \wedge e(X, Y) \Rightarrow rare(Y)$. In this case, there is only one size-k subset of C_Γ where the formula α does not hold, so the accuracy is even higher than in the previous example. Yet the adversarial masking process can select evidence consisting of all true positive literals from Γ, i.e. the evidence will consist of the $rare(c_1)$ literal and all the $e/2$ literals from Γ. Then the set of errors that are made when using the formula α will be the same as in Example 3, despite the fact that the rule is almost perfect on Γ.*

Note that in the examples above, we had perfect knowledge of the accuracy of the rule α on the test example (i.e. we knew the value of $Q_{\Gamma,k}(\alpha)$). In practice, this accuracy needs to be estimated from the training example. In such cases, it can thus happen that a rule α has accuracy 1 on the training example Υ, but still produces many errors on $\kappa(\Gamma)$. We will provide PAC-type bounds for this setting with estimated accuracies in Sections 5. First, however, in Section 4.2 and 4.3 we will look at how bounds can be provided on the number of incorrectly derived literals in the case where $Q_{\Gamma,k}(\alpha)$ is known. As the above examples illustrate, to obtain reasonable bounds, we will need to consider forms of inference which are weaker than classical entailment.

4.2 BOUNDED REASONING USING k-ENTAILMENT

We saw that even for formulas which hold for almost all subsets of Γ, the result of using them for inference can be quite disastrous. This was to a large extent due to the fact that we had inference chains involving a large number of domain elements (constants). This observation suggests a natural way to restrict the kinds of inferences that can be made when imperfect rules are involved.

Definition 4 (k-entailment)**.** *Let k be a non-negative integer, $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example, κ be a masking process, and Φ be a set of closed formulas. We say that a ground formula φ is k-entailed by Φ and $\kappa(\Upsilon)$, denoted $\{\kappa(\Upsilon)\} \cup \Phi \models_k \varphi$, if there is a $\mathcal{C}' \subseteq \mathcal{C}$ such that $|\mathcal{C}'| \leq k$, const$(\varphi) \subseteq \mathcal{C}'$, $\{\kappa(\Upsilon)\langle \mathcal{C}' \rangle\} \cup \Phi$ is consistent and $\{\kappa(\Upsilon)\langle \mathcal{C}' \rangle\} \cup \Phi \models \varphi$.*

In other words, a formula ϕ is k-entailed by Υ and Φ if it can be proved using Φ together with a fragment of $\kappa(\Upsilon)$ induced by no more than k constants, with the additional condition that Φ and this fragment are not contradictory.

Example 5. *Let*

$$\begin{aligned}\Upsilon &= (\{fr(alice, bob), sm(alice)\}, \{alice, bob, eve\}) \\ \kappa(\Upsilon) &= fr(alice \wedge bob) \wedge sm(alice) \\ \Phi &= \{\forall X, Y : fr(X, Y) \wedge sm(X) \Rightarrow sm(Y)\}.\end{aligned}$$

Then $\varphi = sm(bob)$ is 2-entailed from $\kappa(\Upsilon)$ and Φ but not 1-entailed.

Note that, in the setting of Example 4, k-entailment would make at most $k-1$ mistakes. However, 2-entailment would already produce many mistakes in the case of Example 3. So there are cases where k-entailment produces fewer errors than classical logic entailment but, quite naturally, also cases where both produce the same number of errors. Importantly, however, for k-entailment, we can obtain non-trivial bounds on the number of errors.

Next we state two lemmas that follow immediatelly from the respective definitions.

Lemma 1. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example, Φ be a set of constant-free formulas and κ be a masking function. Let $\mathcal{C}_k$ be the set of all size-k subsets of $\mathcal{C}$. Let $\mathcal{H}_{\mathcal{X}}$ denote the set of all ground literals which can be derived using k-entailment from $\{\kappa(\Upsilon)\} \cup \Phi$ and only contain constants from $\mathcal{X}$. Then $\mathcal{H}_{\mathcal{C}} = \bigcup_{\mathcal{S} \in \mathcal{C}_l} \mathcal{H}_{\mathcal{S}}$.*

Lemma 2. *When $\Gamma\langle \mathcal{S} \rangle \models \Phi$ then all ground literals that only contain constants from $\mathcal{S}$ and that are entailed by $\{\kappa(\Gamma\langle \mathcal{S} \rangle)\} \cup \Phi$ must be true in $\Gamma\langle \mathcal{S} \rangle$.*

We now provide a bound on the number of ground literals wrongly k-entailed by a given Φ, assuming that we know its accuracy $Q_{\Gamma,k}(\Phi)$ on the example Γ.

Proposition 6. *Let $\Gamma = (\mathcal{A}, \mathcal{C})$ be an example, Φ be a set of constant-free formulas and κ be a masking process. Next let $\mathcal{F}(\Gamma)$ be the set of all ground literals of a predicate p/a, $a \leq k$, which are k-entailed by $\{\kappa(\Gamma)\} \cup \Phi$ but are false in Γ. Then*

$$|\mathcal{F}(\Gamma)| \leq (1 - Q_{\Gamma,k}(\Phi))|\mathcal{C}|^k k^a.$$

Proof. First, we note that the number of size-k subsets is bounded by $|\mathcal{C}|^k$ and the number of different ground p/a atoms in each of these subsets is k^a. It follows from Lemma 2 and Lemma 1 that for any literal $\delta \in \mathcal{F}$ there must be a size-k set $\mathcal{S} \subseteq \mathcal{C}$ such that $\Gamma\langle \mathcal{S} \rangle \not\models \Phi$. The number of all such $\mathcal{S}$'s that satisfy $\Gamma\langle \mathcal{S} \rangle \not\models \Phi$ is bounded by $(1 - Q_{\Gamma,k}(\Phi))|\mathcal{C}|^k$. Hence, we have $|\mathcal{F}(\Gamma)| \leq (1 - Q_{\Gamma,k}(\Phi))|\mathcal{C}|^k k^a$. $\square$

We can notice that when we increase the domain size $|\mathcal{C}|$, keeping $Q_{\Gamma,k}(\Phi)$ fixed and non-zero, the bound eventually becomes vacuous for predicates whose arity a is strictly smaller than k. This is because the number of all ground literals grows only as $|\mathcal{C}|^a$ whereas the bound grows as $|\mathcal{C}|^k$. However, if $a = k$, the bound stays fixed when we increase the domain size. We will come back to consequences of this fact in Section 6.

4.3 BOUNDED REASONING USING VOTING

To further restrict the set of entailed ground literals, we next introduce *voting entailment*.

Definition 5 (Voting Entailment)**.** *Let k be an integer and $\gamma \in [0; 1]$. Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be an example, Φ be a set of constant-free formulas, and κ be a masking process. A ground literal l of arity a, $a \leq k$, is said to be entailed from Φ and $\kappa(\Upsilon)$ by voting with parameters k and γ if there are at least $\max\{1, \gamma \cdot |\mathcal{C}|^{k-a}\}$ size-k sets $\mathcal{S} \subseteq \mathcal{C}$ such that l is k-entailed by $\kappa(\Upsilon)\langle \mathcal{S} \rangle$.*

The next example illustrates the use of voting entailment.

Example 7. *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$, where $\mathcal{C} = \{alice, bob, eve\}$, and let $\kappa(\Upsilon) = fr(alice, bob) \wedge fr(eve, bob) \wedge sm(eve)$. Next, let $\Phi = \{\forall X, Y : fr(X, Y) \wedge sm(X) \Rightarrow sm(Y)\}$. Then $sm(bob)$ is entailed from Φ and $\kappa(\Upsilon)$ by voting with the parameters $k = 2$ and $\gamma = 2/3$, as $\gamma \cdot |\mathcal{C}|^{k-a} = 2/3 \cdot 3^{2-1} = 2$ and there are two size-2 subsets of $\mathcal{C}$ that 2-entail $sm(bob)$.*

We now show how the bound from Proposition 6 can be strengthened in the case of voting entailment.

Proposition 8. *Let k be an integer and $\gamma \in [0; 1]$. Let $\Gamma = (\mathcal{A}, \mathcal{C})$ be an example, Φ be a set of constant-free formulas, and κ be a masking process. Let $\mathcal{F}(\Gamma)$ be the set of all ground literals of a predicate p/a, $a \leq k$, that are entailed by voting from $\{\kappa(\Gamma)\} \cup \Phi$ with parameters k and γ but are false in Γ. If $\gamma \cdot |\mathcal{C}|^{k-a} \geq 1$ then*

$$|\mathcal{F}(\Gamma)| \leq (1 - Q_{\Gamma,k}(\Phi)) \frac{|\mathcal{C}|^a k^a}{\gamma}$$

and otherwise

$$|\mathcal{F}(\Gamma)| \leq (1 - Q_{\Gamma,k}(\Phi)) |\mathcal{C}|^k k^a.$$

Proof. First we define the number of "votes" for a ground literal l as

$$\#_{\kappa(\Gamma),\Phi}(l) = |\{\mathcal{S} \subseteq \mathcal{C} \,||\mathcal{S}| = k, \{\kappa(\Gamma)\langle \mathcal{S} \rangle\} \cup \Phi \models_k l\}|.$$

Let L be the set of all ground p/a literals l such that $\Gamma \models \neg l$. Then, since any size-k subset of $\mathcal{C}$ can only contribute k^a votes to literals based on the predicate p/a, we have

$$\sum_{l \in L} \#_{\kappa(\Gamma),\Phi}(l) \leq (1 - Q_{\Gamma,k}(\Phi)) \, |\mathcal{C}|^k k^a.$$

Hence $|\mathcal{F}(\Gamma)| \leq \frac{(1-Q_{\Gamma,k}(\Phi))|\mathcal{C}|^k k^a}{\max\{1, \gamma \cdot |\mathcal{C}|^{k-a}\}}$. If $\gamma \cdot |\mathcal{C}|^{k-a} \geq 1$ then $|\mathcal{F}(\Gamma)| \leq (1 - Q_{\Gamma,k}(\Phi)) \frac{|\mathcal{C}|^a k^a}{\gamma}$. The case when $\gamma \cdot |\mathcal{C}|^{k-a} < 1$ follows from Theorem 6. □

Unlike for k-entailment, the fraction of "wrong" ground p/a literals entailed by voting entailment does not grow with an increasing domain size as long as $\gamma \cdot |\mathcal{C}|^{k-a} \geq 1$.

5 PROBABILISTIC BOUNDS

We now turn to the setting where the accuracy of the formulas needs to be estimated from a training example Υ. More generally, we also cover the case where the formulas themselves are learned from the training example. In such cases, to account for over-fitting, we need to consider the (size of the) hypothesis class that was used for learning these formulas. Specifically, we prove probabilistic bounds for variants of the following learning problem. We are given a hypothesis set $\mathcal{H}$ of constant-free theories, and we want to compute bounds on the number of incorrectly predicted literals which simultaneously hold for all $\Phi \in \mathcal{H}$ (as a function of $Q_{\Upsilon,k}(\Phi)$) with probability at least $1 - \delta$, where δ is a confidence parameter. Note that the case where the theory Φ is given, rather than learned, corresponds to $\mathcal{H} = \{\Phi\}$.

We start by proving general concentration inequalities in Section 5.1 which we then use to prove bounds for k-entailment. These bounds are studied for the realizable case in Section 5.2 and for the general case in Section 5.3. Bounds for voting entailment are studied in Section 5.4

5.1 CONCENTRATION INEQUALITIES

We will need to bound the difference between the "accuracy" of given sets of logic formulas Φ on the training sample Υ and their accuracy on a test sample Γ (i.e. the difference between $Q_{\Upsilon,k}(\Phi)$ and $Q_{\Gamma,k}(\Phi)$). To prove the concentration inequalities in this section, we will utilize the following lemma.

Lemma 3 (Kuželka et al. [12]). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be an example. Let $0 \leq n \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathbf{X} = (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{\lfloor \frac{n}{k} \rfloor})$ 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 $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$. Finally, let $\mathcal{I}' = \{1, 2, \ldots, |\mathcal{C}_\aleph|\}$ and let $\mathbf{Y} = (\mathcal{S}'_1, \mathcal{S}'_2, \ldots, \mathcal{S}'_{\lfloor \frac{n}{k} \rfloor})$ be a vector sampled by the following process:*

1. *Sample subsets $\mathcal{I}'_1, \ldots, \mathcal{I}'_{\lfloor \frac{n}{k} \rfloor}$ of size k from $\mathcal{I}'$.*
2. *Sample an injective function $g : \bigcup_{i=1}^{\lfloor n/k \rfloor} \mathcal{I}'_i \rightarrow \mathcal{C}_\Upsilon$ uniformly from all such functions.*
3. *Define $\mathcal{S}'_i = g(\mathcal{I}'_i)$ for all $0 \leq i \leq \lfloor \frac{n}{k} \rfloor$.*

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

The next example illustrates the intuition behind the proof of this lemma, which can be found in [12].

Example 9. *Let $\mathcal{C}_\aleph = \{1, 2, \ldots, 10^6\}$. Let us sample $\lfloor m/k \rfloor$ size-k subsets of $\mathcal{C}_\aleph$ uniformly. If this was the process that generates the data from which we estimate parameters, we could readily apply Hoeffding's inequality to get the confidence bounds. However, in typical SRL settings (e.g. with MLNs), we are given a complete example on some set of constants (objects), rather than a set of small sampled fragments. So we instead need to assume that the whole training example is sampled at once, uniformly from all size-m subsets of $\mathcal{C}_\aleph$. However, when we then estimate the probabilities of formulas from this example, we cannot use Hoeffding's bound or any other bound expecting independent samples. What we can do[2] is to mimic sampling from $\mathcal{C}_\aleph$ by sampling from an auxiliary set of constants of the same size as $\mathcal{C}_\aleph$ and then specialising these constants to constants from a sampled size-m subset. Hence the first $\lfloor m/k \rfloor$ sampled sets will be distributed exactly as the first $\lfloor m/k \rfloor$ subsets sampled i.i.d. directly from $\mathcal{C}_\aleph$.*

Lemma 3 was used in [12] to prove a bound on expected error. Here we extend that result and use Lemma 3 to prove the concentration inequalities stated in the next two theorems.

Theorem 10. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be an example 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 $\widehat{A}_\Upsilon = Q_{\Upsilon,k}(\alpha)$ and let $A_\aleph = Q_{\aleph,k}(\alpha)$. Then we have $P[\widehat{A}_\Upsilon - A_\aleph \geq \varepsilon] \leq \exp\left(-2\left\lfloor \frac{n}{k} \right\rfloor \varepsilon^2\right)$, $P[A_\aleph - \widehat{A}_\Upsilon \geq \varepsilon] \leq \exp\left(-2\left\lfloor \frac{n}{k} \right\rfloor \varepsilon^2\right)$, and $P\left[\left|\widehat{A}_\Upsilon - A_\aleph\right| \geq \varepsilon\right] \leq 2\exp\left(-2\left\lfloor \frac{n}{k} \right\rfloor \varepsilon^2\right)$.*

Proof. First we define an auxiliary estimator $\widetilde{A}_\Upsilon^{(q)}$. Let $\mathbf{Y}^{(q)}$ be a vector of $\lfloor n/k \rfloor \cdot q$ size-k subsets of $\mathcal{C}_\Upsilon$ where

[2] Note that we do not need to do this in practice which will follow from Theorem 10; we only need this mimicking process to prove that theorem.

the subsets of $\mathcal{C}_\Upsilon$ in each of the q non-overlapping size-$\lfloor n/k \rfloor$ segments $\mathbf{Y}_1^{(q)}, \mathbf{Y}_2^{(q)}, \ldots, \mathbf{Y}_q^{(q)}$ of $\mathbf{Y}^{(q)}$ are sampled in the same way as the elements of the vector $\mathbf{Y}$ in Lemma 3, all with the same $\mathcal{C}_\Upsilon$ (i.e. $\mathbf{Y}^{(q)}$ is the concatenation of the vectors $\mathbf{Y}_1^{(q)}, \mathbf{Y}_2^{(q)}, \ldots, \mathbf{Y}_q^{(q)}$). Let us define $\widetilde{A}_\Upsilon^{(q)} = \frac{1}{q \cdot \lfloor n/k \rfloor} \sum_{\mathcal{S} \in \mathbf{Y}^{(q)}} \mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha)$. We can rewrite $\widetilde{A}_\Upsilon^{(q)}$ as $\widetilde{A}_\Upsilon^{(q)} = \frac{1}{q} \sum_{i=1}^{q} \frac{1}{\lfloor n/k \rfloor} \sum_{\mathcal{S} \in \mathbf{Y}_i^{(q)}} \mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha)$.

Then we can use the following trick (Hoeffding [9], Section 5) based on application of Jensen's inequality and Markov's inequality: If $T = a_1 \cdot T_1 + a_2 \cdot T_2 + \cdots + a_q \cdot T_n$, where $a_i \geq 0$ and $\sum_{i=1}^{q} a_i = 1$, then, for any $h > 0$, $P[T \geq \varepsilon] \leq \sum_{i=1}^{n} a_i \cdot \mathbb{E}\left[\exp\left(h(T_i - \varepsilon)\right)\right]$. Note that the T_i's do not have to be independent. Next, using Hoeffding's lemma (Lemma 1 in [9]), if $a_i = 1/q$ and each of the terms T_i is a sum of independent random zero-mean variables $X_j^{(i)}$ such that $P[a \leq X_j^{(i)} \leq b] = 1$ and $b - a \leq 1$, then we get:

$$
\begin{aligned}
P[T \geq \varepsilon] &\leq \sum_{i=1}^{q} \frac{1}{q} \cdot \mathbb{E}\left[\exp\left(h(T_i - \varepsilon)\right)\right] \\
&\leq e^{-h\varepsilon} \exp\left(\frac{m \cdot h^2}{8}\right) = \exp\left(-h\varepsilon + \frac{m \cdot h^2}{8}\right)
\end{aligned}
$$

where m denotes the number of summands of T_i (which, in our case, is the same for all T_i's). Note that this function achieves its minimum at $h = \frac{4\varepsilon}{m}$. We set $T_i := \sum_{\mathcal{S} \in \mathbf{Y}_i^{(q)}} \left(\mathbb{1}(\Upsilon\langle \mathcal{S} \rangle \models \alpha) - A_\aleph\right)$ (note that $\mathbb{E}[T_i] = 0$ and $m = \lfloor n/k \rfloor$). Thus, we get $P[\left\lfloor \frac{n}{k} \right\rfloor \cdot (\widetilde{A}_\Upsilon^{(q)} - A_\aleph) \geq \varepsilon] \leq \exp\left(-2\varepsilon^2 / \left\lfloor \frac{n}{k} \right\rfloor\right)$, and finally

$$
P[\widetilde{A}_\Upsilon^{(q)} - A_\aleph \geq \varepsilon] \leq \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor \varepsilon^2\right),
$$

symmetrically also $P\left[A_\aleph - \widetilde{A}_\Upsilon^{(q)} \geq \varepsilon\right] \leq \exp\left(-2 \left\lfloor \frac{n}{k} \right\rfloor \varepsilon^2\right)$, and, using union bound, we get

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

It follows from the strong law of large numbers (which holds for any Υ) that $P[\lim_{q \to \infty} \widetilde{A}_\Upsilon^{(q)} = \widehat{A}_\Upsilon] = 1$. Since q was arbitrary, the statement of the proposition follows. □

As the next theorem shows, the above result can be generalized to the case where we need to bound the difference between the estimations obtained from two samples.

Theorem 11. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be an example and let $0 \leq n, u \leq |\mathcal{C}_\aleph|$ and $0 \leq k \leq n$ be integers. Let $\mathcal{C}_\Upsilon$ and $\mathcal{C}_\Gamma$ be sampled uniformly from all size-n and size-u subsets of $\mathcal{C}_\aleph$ and let $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle, \Gamma = \aleph\langle \mathcal{C}_\Gamma \rangle$. Let α be a closed and constant-free formula. Let $\widehat{A}_\Upsilon = Q_{\Upsilon,k}(\alpha)$, $\widehat{A}_\Gamma = Q_{\Gamma,k}(\alpha)$, and let $A_\aleph = Q_{\aleph,k}(\alpha)$. Then we have $P[\widehat{A}_\Upsilon - \widehat{A}_\Gamma \geq \varepsilon] \leq \exp\left(\frac{-2\varepsilon^2}{1/\lfloor n/k \rfloor + 1/\lfloor u/k \rfloor}\right)$, and $P\left[\left|\widehat{A}_\Upsilon - \widehat{A}_\Gamma\right| \geq \varepsilon\right] \leq 2 \exp\left(\frac{-2\varepsilon^2}{1/\lfloor n/k \rfloor + 1/\lfloor u/k \rfloor}\right)$.*

Proof. See the appendix. □

We note that the concentration inequality derived in Theorem 10 improves upon a concentration inequality derived in [17] (Chapter 10) that contains n/k^2 (in our notation) instead of $\lfloor n/k \rfloor$ in the exponential.[3]

Next we prove an inequality for the special case where the probability of a formula α on Υ is 0. Since we can also take negations of formulas, this theorem will be useful to prove bounds for formulas that are perfectly accurate on training data. As the following theorem shows, in this case we obtain stronger guarantees, where we have ε instead of ε^2 in the exponential.

Theorem 12. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be an example 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 $\widehat{A}_\Upsilon = Q_{\Upsilon,k}(\alpha)$ and let $A_\aleph = Q_{\aleph,k}(\alpha) \geq \varepsilon$. Then we have*

$$
P\left[\widehat{A}_\Upsilon = 0\right] \leq \exp\left(-\lfloor n/k \rfloor \varepsilon\right).
$$

Proof. Let $\mathbf{Y}$ be sampled as in Lemma 3 (i.e. $\mathbf{Y}$ is sampled only using Υ and not directly $\aleph$). Then using Lemma 3 we know that the elements of $\mathbf{Y}$ are distributed like $\lfloor n/k \rfloor$ independent samples (size-k subsets) from $C_\aleph$. Hence we can bound the probability $P[A_\Upsilon = 0] \leq (1 - \varepsilon)^{\lfloor n/k \rfloor} \leq \exp\left(-\lfloor n/k \rfloor \varepsilon\right)$. Obviously, adding the rest of the information from size-k subsets of C_Υ that are not contained in $\mathbf{Y}$ cannot increase the bound. □

5.2 ZERO TRAINING ERROR CASE

We start by proving a bound for the realizable (i.e. zero training error) case.

Theorem 13. *Let $\aleph$, Υ, Γ, n, u and κ be as in Definition 3 (i.e. Υ and Γ are sampled from $\aleph$ and n, u are sizes of Υ's and Γ's domains). Let $\mathcal{H}$ be a finite hypothesis class of constant-free formulas. Let $\mathcal{F}(\Gamma, \Phi)$ denote the set of all ground literals of a predicate p/a that are k-entailed by $\{\kappa(\Gamma)\} \cup \Phi$ but are false in Γ.*[4] *With probability at*

[3] This is essentially due to the fact that we use Hoeffding's decomposition whereas Lovasz relies on Azuma's inequality, leading to a looser bound compared to our bound.

[4] Note that here, as well as in the rest of the theorems in the paper, $\mathcal{F}(\Gamma, \Phi)$ is a set-valued random variable.

least $1-\delta$*, the following holds for all* $\Phi \in \mathcal{H}$ *that satisfy* $Q_{\Upsilon,k}(\Phi) = 1$*:*

$$\mathbb{E}\left[|\mathcal{F}(\Gamma,\Phi)|\right] \leq \frac{\ln|\mathcal{H}| + \ln 1/\delta}{\lfloor n/k \rfloor} u^k k^a.$$

Proof. It follows from the linearity of expectation and from Proposition 6 that, for any Φ, $\mathbb{E}\left[|\mathcal{F}(\Gamma,\Phi)|\right] \leq (1 - Q_{\aleph,k}(\Phi))u^k k^a$. Next, it follows from Theorem 12 and from the union bound taken over all $\Phi \in \mathcal{H}$ that the probability that there exists $\Phi \in \mathcal{H}$ such that $Q_{\Upsilon,k}(\Phi) = 1$ and $\varepsilon \leq 1 - Q_{\aleph,k}(\Phi)$ is at most $|\mathcal{H}| \cdot \exp\left(-\lfloor n/k \rfloor \varepsilon\right)$. If $\varepsilon \geq \frac{\ln|\mathcal{H}| + \ln 1/\delta}{\lfloor n/k \rfloor}$ then $|\mathcal{H}| \cdot \exp\left(-\lfloor n/k \rfloor \varepsilon\right) \leq \delta$. Hence, with probability at least $1-\delta$, the following holds for all $\Phi \in \mathcal{H}$ such that $Q_{\Upsilon,k}(\Phi) = 1$: $\mathbb{E}\left[|\mathcal{F}(\Gamma,\Phi)|\right] \leq \frac{\ln|\mathcal{H}| + \ln 1/\delta}{\lfloor n/k \rfloor} u^k k^a$. □

5.3 GENERAL CASE

Next we prove a bound for the general case when the training error is non-zero.

Theorem 14. *Let* $\aleph$, Υ, Γ, n, u *and* κ *be as in Definition 3 (i.e.* Υ *and* Γ *are sampled from* $\aleph$ *and* n, u *are sizes of* Υ*'s and* Γ*'s domains). Let* $\mathcal{H}$ *be a finite hypothesis class of constant-free formulas. Let* $\mathcal{F}(\Gamma,\Phi)$ *denote the set of all ground literals of a predicate* p/a *that are* k*-entailed by* $\{\kappa(\Gamma)\} \cup \Phi$ *but are false in* Γ*. With probability at least* $1-\delta$*, for all* $\Phi \in \mathcal{H}$*:*

$$\mathbb{E}\left[|\mathcal{F}(\Gamma,\Phi)|\right] \leq \left(1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{\ln\left(\frac{|\mathcal{H}|}{\delta}\right)}{2\lfloor n/k \rfloor}}\right) u^k k^a.$$

Proof. First, as in the proof of Theorem 13, we find that, for any $\Phi \in \mathcal{H}$, $\mathbb{E}\left[|\mathcal{F}(\Gamma)|\right] \leq (1 - Q_{\aleph,k}(\Phi))u^k k^a$. Next, it follows from Theorem 10 and from union bound that $P\left[\exists \Phi \in \mathcal{H} : Q_{\Upsilon,k}(\Phi) - Q_{\aleph,k}(\Phi) \geq \varepsilon\right] \leq |\mathcal{H}| \exp\left(-2\lfloor n/k \rfloor \varepsilon^2\right)$. It follows that

$$P\left[\exists \Phi \in \mathcal{H} : Q_{\Upsilon,k}(\Phi) \geq Q_{\aleph,k}(\alpha) + \sqrt{\frac{\ln\left(|\mathcal{H}|/\delta\right)}{2\lfloor n/k \rfloor}}\right] \leq \delta.$$

The theorem then follows straightforwardly from the above and from Proposition 6. □

The previous two theorems provided bounds on the expected number of errors on the sampled test examples. The next theorem is different in that it provides a bound on the actual number of errors.

Theorem 15. *Let* $\aleph$, Υ, Γ, *and* κ *be as in Definition 3 (i.e.* Υ *and* Γ *are sampled from* $\aleph$ *and* n, u *are sizes of* Υ*'s and* Γ*'s domains). Let* $\mathcal{H}$ *be a finite hypothesis class of constant-free formulas. Let* $\mathcal{F}(\Gamma,\Phi)$ *denote the set of all ground literals of a predicate* p/a *that are* k*-entailed by* $\{\kappa(\Gamma)\} \cup \Phi$ *but are false in* Γ*. With probability at least* $1-\delta$*, for all* $\Phi \in \mathcal{H}$ *:*

$$|\mathcal{F}(\Gamma,\Phi)| \leq \left(1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{(\lfloor n/k \rfloor + \lfloor u/k \rfloor)\ln\left(2|\mathcal{H}|/\delta\right)}{2\lfloor n/k \rfloor \lfloor u/k \rfloor}}\right) u^k k^a$$
$$\leq \left(1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{\ln\left(2|\mathcal{H}|/\delta\right)}{\min(\lfloor n/k \rfloor, \lfloor u/k \rfloor)}}\right) u^k k^a.$$

Proof. Let us denote $\widehat{A} = Q_{\Upsilon,k}(\Phi)$, $\widehat{B} = Q_{\Gamma,k}(\Phi)$. Using Theorem 11 and the union bound over $\Phi \in \mathcal{H}$, we get

$$P[\exists \Phi \in \mathcal{H} : |\widehat{A} - \widehat{B}| \geq \varepsilon] \leq 2|\mathcal{H}| \exp\left(\frac{-2\varepsilon^2 \lfloor n/k \rfloor \lfloor u/k \rfloor}{\lfloor n/k \rfloor + \lfloor u/k \rfloor}\right).$$

Solving the above for ε that achieves the $1-\delta$ bound, we obtain that, with probability at least $1-\delta$, we have for all $\Phi \in \mathcal{H}$: $|\widehat{A} - \widehat{B}| \leq \sqrt{\frac{(\lfloor n/k \rfloor + \lfloor u/k \rfloor)\ln(2|\mathcal{H}|/\delta)}{2\lfloor n/k \rfloor \lfloor u/k \rfloor}}$. Hence, with probability at least $1-\delta$, for all $\Phi \in \mathcal{H}$ it holds $1 - Q_{\Gamma,k}(\Phi) \leq 1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{(\lfloor n/k \rfloor + \lfloor u/k \rfloor)\ln(2|\mathcal{H}|/\delta)}{2\lfloor n/k \rfloor \lfloor u/k \rfloor}}$. The validity of the theorem then follows from the above and from Proposition 6 and the fact that $\frac{ab}{a+b} \geq \frac{\min(a,b)}{2}$ for any nonnegative a and b. □

5.4 BOUNDS FOR VOTING ENTAILMENT

Next we prove a bound for voting entailment, which, unsurprisingly, is tighter than the respective bound for k-entailment.

Theorem 16. *Let* k *be an integer and* $\gamma \in [0;1]$*. Let further* $\aleph$, Υ, Γ *and* κ *be as in Definition 3 (i.e.* Υ *and* Γ *are sampled from* $\aleph$ *and* n, u *are sizes of* Υ*'s and* Γ*'s domains). Let* $\mathcal{H}$ *be a finite hypothesis class of constant-free formulas. Let* $\mathcal{F}(\Gamma,\Phi)$ *denote the set of all ground literals of a predicate* p/a *that are entailed by voting from* $\{\kappa(\Gamma)\} \cup \Phi$ *with parameters* k *and* γ *but are false in* Γ*. Then, with probability at least* $1-\delta$*, for all* $\Phi \in \mathcal{H}$*:*

$$|\mathcal{F}(\Gamma)| \leq \left(1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{\ln\left(2|\mathcal{H}|/\delta\right)}{\min\left\{\lfloor u/k \rfloor, \lfloor n/k \rfloor\right\}}}\right) \frac{u^a k^a}{\gamma}.$$

Proof. This follows from the same reasoning as in the proof of Theorem 15, which gives us the bound on the difference of $Q_{\Upsilon,k}(\Phi)$ and $Q_{\Gamma,k}(\Phi)$, combined with Theorem 8. □

Remark 17. *The fraction of "wrong" ground p/a literals does not grow with increasing test-set size (u), since, by rewriting the bound from Theorem 16, we get, with probability at least $1-\delta$, for all $\Phi \in \mathcal{H}$:*

$$\frac{|\mathcal{F}(\Gamma)|}{u^a} \leq \left(1 - Q_{\Upsilon,k}(\Phi) + \sqrt{\frac{\ln(2|\mathcal{H}|/\delta)}{\min\{\lfloor u/k \rfloor, \lfloor n/k \rfloor\}}}\right) \frac{k^a}{\gamma}.$$

We note here that one can also easily obtain counterparts of Theorems 13 and 14 for voting entailment.

6 SUMMARY OF RESULTS

In this section we discuss positive and negative results that follow from the theorems presented in the preceding sections. Here, bounds are considered vacuous if they are not lower than the total number of ground literals. We first focus on k-entailment in Sections 6.1–6.3, and then discuss the results for voting entailment in Section 6.4. Finally, we also make a connection to MAP-entailment in Section 6.5.

6.1 SMALL TEST EXAMPLES

One case where we have non-vacuous bounds for the *expected* number of incorrectly predicted literals with k-entailment is when the domain of the test examples Γ is small. Naturally a necessary condition is also that the given (or learned) theory Φ is sufficiently accurate. The only way to be confident that Φ is indeed sufficiently accurate, given that this accuracy needs to be estimated, is by estimating it on a sufficiently large training example. This is essentially what Theorems 13 and 14 imply.

Interestingly, this finding agrees with some experimental observations in the literature. For instance, it has been observed in [14] that classical reasoning in a relational setting close to ours worked well for small-size test-set evidence but was not competitive with other methods for larger evidence sizes. The analysis in the present paper thus sheds light on experimental observations like these.

Note that the bounds from Theorems 13 and 14 are for the *expected* value of the number of errors. Bounds on the actual number of errors are provided in Theorem 15. In this case, to obtain non-vacuous bounds, we also need to require that the domain of the test example Γ be sufficiently large. This is not unexpected, however, as it is a known property of statistical bounds for transductive settings (see e.g., [21]) that the size of the test set affects confidence bounds, similarly to how the size of the Γ's domain affects the bound in Theorem 15.

6.2 PREDICATES OF ARITY K

Another case where we have non-vacuous bounds for k-entailment is when the arity of the predicted literals is equal to the parameter k. In this case both the bounds for the expected error and for the actual error $|\mathcal{F}(\Gamma,\Phi)|$ are non-vacuous. This means that our results cover important special cases. One such special case is classical attribute-value learning when $k = 1$ and we represent attributes by unary predicates. Another case is link prediction when $k = 2$ and higher-arity versions thereof. In link prediction, we have rules such as, for instance, $\forall X, Y : \mathit{CoensFan}(X) \wedge \mathit{CoensFilm}(Y) \Rightarrow \mathit{likes}(X, Y)$.

6.3 REALIZABLE SETTING

We can get stronger guarantees when the given (or learned) theory Φ has zero training error. Keeping the fraction of the domain-sizes $|\mathcal{C}_\Gamma|^{k-a}/|\mathcal{C}_\Upsilon|$ small, Theorem 13 implies non-vacuous bounds for predicates of arity a for any size of the domain of Γ. Intuitively, this means that we can use theories that are completely accurate on training data for inference using k-entailment. However, the required size of the domain of the training example Υ, to guarantee that we will not produce too many errors, grows exponentially with k (for a fixed arity a) and polynomially with $|\mathcal{C}_\Gamma|$.

6.4 VOTING

When using voting entailment, we can always obtain non-trivial bounds by making γ large; obviously this comes at the price of making the inferences more cautious. Voting entailment is a natural inference method in domains where one proof is not enough, i.e. where the support from several proofs is needed before we can be sufficiently confident in the conclusion; an example of such a domain is the well-known *smokers* domain, where knowing that one friend smokes does not provide enough evidence to conclude that somebody smokes; only if we have evidence of several smoker friends is the conclusion warranted that this person smokes.

6.5 RELATIONSHIP TO MAP INFERENCE

A popular approach to collective classification in relational domains is MAP-inference in Markov logic networks. Therefore a natural question is how this approach performs in our setting. Perhaps surprisingly, it might produce as many errors as classical logic reasoning in the examples from Section 4.1, if the Markov logic network contains the same rules, all with positive weights, as we had in these examples. This is because MAP-inference will predict the same literals as classical logical inference when the rules from the Markov logic network are consistent with the given evidence. Thus, we can see that our guarantees for both k-entailment and voting entailment are better than guarantees one could get for MAP-inference. This is also in agreement with the well-known observations that, for instance, in the smokers domain, MAP inference often predicts everyone to be a smoker or everyone to be a non-smoker if there is only a small amount of evidence.

7 RELATED WORK

Our main inspiration comes from the works on PAC-semantics by Valiant [23] and Juba [10]. Our work differs mainly in the fact that we have one large relational structure $\aleph$, and a training example Υ and a test example Γ, both sampled from $\aleph$, whereas it is assumed in these existing approaches that learning examples are sampled i.i.d. from some distribution. This has two important consequences. First, they could use statistical techniques developed for i.i.d. data whereas we had to first derive concentration inequalities for sampling without replacement in the relational setting. Second, since they only needed to bound the error on the independently sampled examples, they did not have to consider the number of incorrectly inferred facts. In contrast, in the relational setting that we considered here, the number of errors made on one relational example is the quantity that needs to be bounded. It follows that completely different techniques are needed in our case. Another difference is that, in their case, the training examples are also masked. In principle, we could modify our results to accommodate for masked examples by replacing "accurate" formulas by sufficiently-often "witnessed" formulas (see [10] for a definition).

Dhurandhar and Dobra [5] derived Hoeffding-type inequalities for classifiers trained with relational data, but these inequalities, which are based on the restriction on the independent interactions of data points, cannot be applied to solve the problems considered in the present paper. Certain other statistical properties of learning have also been studied for SRL models. For instance, Xiang and Neville [25] studied consistency of estimation. However, guaranteeing convergence to the correct distribution does not mean that the model would not generate many errors when used, e.g., for MAP-inference. In [26], they further studied errors in label propagation in collective classification. In their setting, however, the relational graph is fixed and one only predicts labels of vertices exploiting the relational structure for making the predictions. Here we also note that it is not always possible or desirable in practice to sample sets of domain elements uniformly as we assumed to be the case in our analysis. Other sampling designs for relational data were studied, e.g. in [1]. A study of PAC guarantees for such other sampling designs is left as a topic for future work.

There have also been works studying restricted forms of inference in a purely logical context, e.g. [6]. It is an interesting question for future work to find out which existing restricted inference systems would lead to non-vacuous error bounds in the relational setting.

8 CONCLUSIONS

We have studied the problem of predicting plausible missing facts in relational data, given a set of imperfect logical rules, in a PAC reasoning setting. As for the considered inference methods, one of our main objectives was for the inference methods to stay close to classical logic. The first inference method, k-entailment, is a restricted form of classical logic inference and hence satisfies this objective. The second inference method, voting entailment, is based on a form of voting that combines results from inferences made by k-entailment on subsets of the relational data. Importantly, the voting is not weighted which makes voting entailment easier to understand. We were able to obtain non-trivial bounds for the number of literals incorrectly predicted by a learned (or given) theory for both k-entailment and voting entailment. Probably the most useful results of our analysis lie in the identification of cases where the bounds for learning and reasoning in relational data are non-vacuous, which we discussed in detail in Section 6.

There are many interesting directions in which one could extend the results presented in this paper. For instance, as practical means to improve the explainability of inferences made by voting entailment, we could first find representatives of isomorphism classes of "proofs" that are aggregated by voting entailment, and only show these to the user. Another direction is to extend the notion of implicit learning from [10] into the relational setting. It would also be interesting to exploit explicit sparsity constraints and to study other sampling designs, although that might also turn out to be analytically less tractable than the setting considered in the present paper. Finally, although all bounds presented in this paper assume finite hypothesis classes, we note that it is also possible to extend our results to infinite hypothesis classes [13].

Acknowledgments

OK's work was partially supported by the Research Foundation - Flanders (project G.0428.15). SS is supported by ERC Starting Grant 637277. JD is partially supported by the KU Leuven Research Fund (C14/17/070,C22/15/015,C32/17/036), and FWO-Vlaanderen (SBO-150033).

References

[1] Nesreen K Ahmed, Jennifer Neville, and Ramana Rao Kompella. Network sampling designs for relational classification. In *ICWSM*, 2012.

[2] Leila Amgoud. Postulates for logic-based argumentation systems. *International Journal of Approximate Reasoning*, 55(9):2028–2048, 2014.

[3] Newton CA Da Costa et al. On the theory of inconsistent formal systems. *Notre dame journal of formal logic*, 15(4):497–510, 1974.

[4] Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. ProbLog: a probabilistic prolog and its application in link discovery. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2468–2473. Morgan Kaufmann Publishers Inc., 2007.

[5] Amit Dhurandhar and Alin Dobra. Distribution-free bounds for relational classification. *Knowledge and information systems*, 31(1):55–78, 2012.

[6] Marcello DAgostino, Marcelo Finger, and Dov Gabbay. Semantics and proof-theory of depth bounded boolean logics. *Theoretical Computer Science*, 480:43–68, 2013.

[7] Morten Elvang-Gøransson and Anthony Hunter. Argumentative logics: Reasoning with classically inconsistent information. *Data & Knowledge Engineering*, 16(2):125–145, 1995.

[8] P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

[9] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[10] Brendan Juba. Implicit learning of common sense for reasoning. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 939–946, 2013.

[11] Sébastien Konieczny and Ramón Pino Pérez. Merging information under constraints: a logical framework. *Journal of Logic and computation*, 12(5):773–808, 2002.

[12] Ondřej Kuželka, Yuyi Wang, Jesse Davis, and Steven Schockaert. Relational marginal problems: Theory and estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. AAAI Press, 2018.

[13] Ondřej Kuželka, Yuyi Wang, and Steven Schockaert. VC-dimension based generalization bounds for relational learning. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018.

[14] Ondřej Kuželka, Jesse Davis, and Steven Schockaert. Induction of interpretable possibilistic logic theories from relational data. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1153–1159. AAAI Press, 2017.

[15] Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 529–539, 2011.

[16] Qiao Liu, Liuyi Jiang, Minghao Han, Yao Liu, and Zhiguang Qin. Hierarchical random walk inference in knowledge graphs. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 445–454, 2016.

[17] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[18] Graham Priest. The logic of paradox. *Journal of Philosophical logic*, 8(1):219–241, 1979.

[19] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[20] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 3791–3803, 2017.

[21] Ilya Tolstikhin and David Lopez-Paz. Minimax lower bounds for realizable transductive classification. *arXiv preprint arXiv:1602.03027*, 2016.

[22] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[23] Leslie G. Valiant. Knowledge infusion. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1546–1551. AAAI Press, 2006.

[24] Gustav Šourek, Suresh Manandhar, Filip Železný, Steven Schockaert, and Ondřej Kuželka. Learning predictive categories using lifted relational neural networks. In *Proceedings of the 26th International Conference on Inductive Logic Programming*, pages 108–119, 2016.

[25] Rongjing Xiang and Jennifer Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 779–788, 2011.

[26] Rongjing Xiang and Jennifer Neville. Understanding propagation error and its effect on collective classification. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 834–843. IEEE, 2011.