VC-Dimension Based Generalization Bounds for Relational Learning

Ondřej Kuželka[1], Yuyi Wang[2], and Steven Schockaert[3]

[1] Department of Computer Science, KU Leuven, Leuven, Belgium
ondrej.kuzelka@kuleuven.be
[2] DISCO Group, ETH Zurich, Zurich, Switzerland
yuwang@ethz.ch
[3] School of Computer Science & Informatics, Cardiff University, Cardiff, UK
SchockaertS1@cardiff.ac.uk

Abstract. In many applications of relational learning, the available data can be seen as a sample from a larger relational structure (e.g. we may be given a small fragment from some social network). In this paper we are particularly concerned with scenarios in which we can assume that (i) the domain elements appearing in the given sample have been uniformly sampled without replacement from the (unknown) full domain and (ii) the sample is complete for these domain elements (i.e. it is the full substructure induced by these elements). Within this setting, we study bounds on the error of sufficient statistics of relational models that are estimated on the available data. As our main result, we prove a bound based on a variant of the Vapnik-Chervonenkis dimension which is suitable for relational data.

1 Introduction

In one of the most common settings in statistical relational learning (SRL), we are given a fragment of a relational structure (i.e. a *training example*) from which we want to learn a model for making predictions about the unseen parts of the structure. For example, the relational structure could correspond to a large social network and the training example to a fragment of the social network specifying the relationships that hold among a small sample of the users, along with their attributes. Clearly, in order to provide any guarantees on the accuracy of these predictions, we need to make (simplifying) assumptions about how the training structures are obtained. In this paper, we follow the setting from [9,8], where it is assumed that these structures are all obtained as fragments induced by domain elements sampled uniformly without replacement.

The specific problem that we consider in this paper is to bound the error that we make when estimating probabilities of first-order theories from the training example, or more specifically, the probability that a first-order theory Φ is satisfied in a small randomly sampled fragment of the relational structure. While this setting has already been studied in [9,8,7], one important remaining problem, which will be the focus of this paper, relates to how the theory Φ is obtained. Typically, Φ is chosen from some hypothesis class, based on the same training example that is used to estimate its probability. The bounds that were derived in [7] for such cases depend on the size of this hypothesis class.

Unfortunately, this can quickly lead to vacuous bounds in many cases. In fact, in some applications, the most natural hypothesis classes are either infinite or so large that they are effectively infinite for all practical purposes. This is the case, for instance, whenever we want to use constructs involving numerical expressions. To address this issue, in this paper we derive bounds which depend on the VC-dimension of the hypothesis class, instead of its size. In this way, we can also obtain, in many cases, tighter bounds than the ones we derived in [7]. To the best of our knowledge, the bounds we introduce in this paper are the first VC-dimension based bounds for relational learning problems.

2 Preliminaries

In this paper we consider function-free language $\mathcal{L}$, which is built from a finite set of constants *Const*, a set of variables *Var* and a set of predicates $\mathit{Rel} = \bigcup_i \mathit{Rel}_i$, where Rel_i contains the predicates of arity i. Throughout this paper we assume that the sets *Const*, *Var* and *Rel* are fixed. For $a_1, ..., a_k \in \mathit{Const} \cup \mathit{Var}$ and $R \in \mathit{Rel}_k$, we call $R(a_1, ..., a_k)$ an *atom*. If $a_1, .., a_k \in \mathit{Const}$, this atom is called *ground*. A *literal* is an atom or its negation. A formula is called *closed* if all variables are bound by a quantifier. Note that although the set *Const* is required to be finite, it can have arbitrary size, so that we could, for instance, represent all 64-bit floating point numbers. From an application point of view, this allows us to consider formulas involving numerical expressions. For example, we could have a predicate *Sum*, whose intended meaning is that $\mathit{Sum}(x, y, z)$ holds iff $z = x + y$ where $+$ represents floating-point addition.

2.1 Relational Learning Setting

Relational examples The learning setting considered in this paper follows the one that was introduced in [9,8]. The central notion is that of a *relational example* (or simply *example* if there is no cause for confusion),which is defined as 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 $\mathcal{C}$. A relational example is intended to provide a complete description of a possible 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 $\mathcal{A}$. For instance, the relational example $(\{\mathit{sm}(alice)\}, \{\mathit{alice}\})$ is different from $(\{\mathit{sm}(alice)\}, \{\mathit{alice}, \mathit{bob}\})$, as in the latter case we know that *bob* does not smoke (i.e. the atom $\mathit{sm}(\mathit{bob})$ is known to be false since it is not specified to be true) whereas in the former case we have no knowledge about *bob*. We denote by $\Omega(\mathcal{L}, k)$ the set of all possible relational examples $\Upsilon = (\mathcal{A}, \mathcal{C})$ where $\mathcal{A}$ only contains ground atoms from $\mathcal{L}$ and $|\mathcal{C}| = k$.

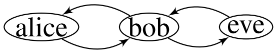
Example 1. Let us assume that the only predicate in $\mathcal{L}$ is $\mathit{sm}/1$ and the only constant in is *alice*. Then $\Omega(\mathcal{L}, 1) = \{(\mathit{sm}(\mathit{alice}), \{\mathit{alice}\}), (\emptyset, \{\mathit{alice}\})\}$.

Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a relational example and $\mathcal{S} \subseteq \mathcal{C}$. The fragment $\Upsilon\langle\mathcal{S}\rangle = (\mathcal{B}, \mathcal{S})$ 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 $\mathcal{S}$.

Example 2. 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 $\Upsilon\langle\{alice, bob\}\rangle = (\{sm(alice), fr(alice, bob), fr(bob, alice)\}, \{alice, bob\})$.

In the considered setting, we are given a single relational example $\Upsilon = (\mathcal{A}, \mathcal{C})$, and this example is assumed to have been sampled from a larger relational example $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$. The intended meaning is that $\aleph$ covers the entire domain which we would like to model and Υ is the fragment of the domain which is known at training time. Throughout this paper, we will assume that $\mathcal{C}_\aleph$ is finite. As in [8,7] we assume that Υ as sampled from $\aleph$ by the following process.

Definition 1 (Sampling from a global example). *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a relational example called the* global example. *Let $n \in \mathbb{N} \setminus \{0\}$ and let Unif$(\mathcal{C}_\aleph, n)$ denote uniform distribution on size-n subsets of $\mathcal{C}_\aleph$. Training relational examples Υ are sampled from the global example $\aleph$ by first sampling $\mathcal{C}_\Upsilon \sim$ Unif$(\mathcal{C}_\aleph, n)$ and defining $\Upsilon = \aleph\langle\mathcal{C}_\Upsilon\rangle$.*

Probabilities of formulas In a given relational example, any closed formula α is classically either true or false. To assign probabilities to formulas in a meaningful way, considering that we typically only have a single relational example available for training, we can consider how often the formula is satisfied in small fragments of the given relational example.

Definition 2 (Probability of a formula [8]). *Let $\Upsilon = (\mathcal{A}, \mathcal{C})$ be a relational example and $k \in \mathbb{N}$. The probability of a closed formula α is defined as follows*[4]*:*

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

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 can straightforwardly be extended to probabilities of sets of formulas (which we will also call *theories* interchangeably): if Φ is a set of formulas, we then have $Q_{\Upsilon,k}(\Phi) = Q_{\Upsilon,k}(\bigwedge \Phi)$ where $\bigwedge \Phi$ denotes the conjunction of all formulas in Φ.

Example 3. 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 consider the following 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 $Q_{\Upsilon,2}(\beta) = 1/3$.

[4] 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.

It is not difficult to check that under the sampling assumption from Definition 1, for any theory Φ it holds that $Q_{\aleph,k}(\Phi) = \mathbb{E}_{\Upsilon}\left[Q_{\Upsilon,k}(\Phi)\right]$ [8].

Representing theories as functions By definition, to compute $Q_{\Upsilon,k}(\Phi)$, we only need to know for which of the elements of $\Omega(\mathcal{L},k)$ it holds that Φ is satisfied. To make this view explicit, we will formulate the results in this paper in terms of functions from $\Omega(\mathcal{L},k)$ to $\{0,1\}$. For a given theory Φ, the associated function f_Φ is defined for $\Gamma \in \Omega(\mathcal{L},k)$ as $f_\Phi(\Gamma) = 1$ if $\Gamma \models \Phi$ and $f_\Phi(\Gamma) = 0$ otherwise. The advantage of this formulation is that our results then directly apply to settings where other representation frameworks than classical logic are used for representing the theory. For example, a theory could be implicitly represented by a neural network with a hard-thresholding output unit. For notational convenience, we also write $\Gamma \models f$ if $f(\Gamma) = 1$. We then naturally extend the definition of $Q_{\Upsilon,k}$ to functions: $Q_{\Upsilon,k}(f) = P_{\mathcal{S}\sim \mathit{Unif}(\mathcal{C},k)}\left[\Upsilon\langle\mathcal{S}\rangle \models f\right] = P_{\mathcal{S}\sim \mathit{Unif}(\mathcal{C},k)}\left[f(\Upsilon\langle\mathcal{S}\rangle) = 1\right].$

2.2 VC-Dimension

The next definition describes the classical notion of VC-dimension [15], specialized to our relational learning setting that is used throughout this paper to measure the complexity of hypothesis classes.

Definition 3 (VC-dimension). *Let k be a positive integer and let $\mathcal{H}$ be a hypothesis class of functions $f : \Omega(\mathcal{L},k) \to \{0,1\}$. Let $\mathcal{X} = \{\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_d\} \subseteq \Omega(\mathcal{L},k)$. We say that $\mathcal{H}$ shatters $\mathcal{X}$ if for every $\mathcal{Y} \subseteq \mathcal{X}$, there is $f \in \mathcal{H}$ such that $f(\Upsilon) = 1$ for all $\Upsilon \in \mathcal{Y}$ and $f(\Upsilon) = 0$ for all $\Upsilon \in \mathcal{X} \setminus \mathcal{Y}$. The VC dimension of $\mathcal{H}$ is the largest integer d such that there exists a subset of $\Omega(\mathcal{L},k)$ with cardinality d that is shattered by $\mathcal{H}$.*

The next definition formalizes what we mean when we say that two functions are equivalent w.r.t. a given global example.

Definition 4. *We say two functions f and g are k-equivalent w.r.t. a global example $\aleph$ if for any size-k set $\mathcal{S}$ it holds that $f(\aleph\langle\mathcal{S}\rangle) = g(\aleph\langle\mathcal{S}\rangle)$.*

Naturally the above two definitions can also be applied to theories, e.g. two theories Φ and Θ are k-equivalent w.r.t. a global example $\aleph$ if their associated functions f_Φ and f_Θ are k-equivalent. The following observation will play an important role in the proofs.

Remark 1 *The maximum number of hypotheses that are mutually non-equivalent w.r.t. a given (finite) global example $\aleph$ is finite.*

A consequence of this observation is that even for infinite hypothesis classes, in principle, there are only finitely many different hypotheses that need to be considered. However, given that we typically do not know the size of the global example, in practice it is not possible to rely on the number of non-equivalent hypotheses to apply the bounds from [7] to infinite hypothesis classes. In contrast, the bounds that we introduce in this paper can still be applied in such cases, as long as the hypothesis class has a finite VC-dimension.

The ability to deal with infinite hypothesis classes makes it possible, for instance, to learn theories based on differentiable architectures [18,12] or based on graph kernels [17].

3 Motivation

The main aim of this paper is to derive bounds on how accurately we can estimate $Q_{\aleph,k}(f)$ from a given training relational example $\varUpsilon$, where f is viewed as a logical formula. The need for such probability estimates naturally arises, among others, in the setting of relational marginal problems, which were studied in [8]. In that setting, we are given a set of formulas $\varTheta = \{\alpha_1, \dots, \alpha_{|\varTheta|}\}$, a set of constants $\mathcal{C}$ and a training relational example $\varUpsilon = (\mathcal{A}_\varUpsilon, \mathcal{C}_\varUpsilon)$. The task is to use the probabilities of $\alpha_1, \dots, \alpha_{|\varTheta|}$ that are estimated from the training relational example $\varUpsilon$ to perform inference on the domain $\mathcal{C}$. Specifically, the task is to find a maximum entropy distribution on the set of all relational examples of the form $\varPsi = (\mathcal{A}_\varPsi, \mathcal{C})$, such that $\mathbb{E}[Q_{\varPsi,k}(\alpha_i)] = \widehat{Q}_{\varUpsilon,k}(\alpha_i)$ for all $\alpha_i \in \varTheta$. Here, $\widehat{Q}_{\varUpsilon,k}(\alpha_i)$ is an estimate of $\mathbb{E}[Q_{\varPsi,k}(\alpha_i)]$ which is based on $Q_{\varUpsilon,k}(\alpha_i)$. If $|\mathcal{C}| \leq |\mathcal{C}_\varUpsilon|$ then this estimate is simply given by $\widehat{Q}_{\varUpsilon,k}(\alpha_i) = Q_{\varUpsilon,k}(\alpha_i)$. In general, however, the value $Q_{\varUpsilon,k}(\alpha_i)$ needs to be adjusted to account for the difference in the size of the training relational example domain $\mathcal{C}_\varUpsilon$ and the domain $\mathcal{C}$ over which we want to perform inference. The resulting distribution is similar to a Markov logic network, and can be used in applications for similar purposes[5]; it is an exponential family distribution of the following form:

$$P(\varPsi) = \frac{1}{Z} \exp\left(\sum_{\alpha_i \in \varTheta} w_i \cdot Q_{\varPsi,k}(\alpha_i) \right).$$

In the case $|\mathcal{C}| = |\mathcal{C}_\varUpsilon|$, the weights w_i can be obtained by solving a maximum likelihood problem which is the dual of the maximum entropy problem. Ideally, we would use $Q_{\aleph,k}(\alpha_i)$ as the estimates of $Q_{\varPsi,k}(\alpha_i)$ in the maximum entropy problems. Since, in reality, we do not have access to $Q_{\aleph,k}(\alpha_i)$, we need to use the estimates based on $Q_{\varUpsilon,k}(\alpha_i)$. The results we present in this paper shed light on the impact of this simplification. We refer the reader to [8] for more details.

Estimates of $Q_{\aleph,k}(f)$ also play a central role in the analysis of PAC-reasoning [6,14] for relational domains as studied in [7]. This analysis also relies on the sampling assumptions from Definition 1. Specifically, in that setting, a training relational example $\varUpsilon$ and a test relational example $\varPsi$ are sampled from $\aleph$ and the learner's task is to find a set of first-order logic formulas that will not produce too many errors on $\varPsi$ when using a restricted form of classical reasoning. To obtain guarantees on the number of literals that are incorrectly inferred using this form or reasoning, we essentially need to bound the difference of $Q_{\varUpsilon,k}(\varPhi)$ and $Q_{\aleph,k}(\varPhi)$ (which allows us to bound the difference with $Q_{\varPsi,k}(\varPhi)$), which is exactly the problem we also study in this paper. In contrast to [7], however, we are interested in bounds that are based on the VC-dimension of the hypothesis space.

[5] The relational marginal problems that we consider in this paper are referred to as Model A in [8]. Another type of relational marginal problems, referred to as Model B in [8], leads to distributions that are exactly Markov logic networks.

4 Summary of the Results

Intuitively, what we need to find is a suitable bound on the quantity $|Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$, i.e. we want to bound the error we make when estimating the overall probability of f (i.e. the value $Q_{\aleph,k}(f)$) from a training fragment of the global example. In most application settings, however, f itself is also chosen using the training relational example Υ, e.g. by choosing the hypothesis f that maximizes $Q_{\Upsilon,k}(f)$ among the functions from some hypothesis class $\mathcal{H}$. This means that we cannot find a suitable bound for $|Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$ without taking the hypothesis class $\mathcal{H}$ into account. The classical solution, which we will also follow, is to instead bound the quantity $\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$. The main result of this paper takes the form of two theorems that provide probabilistic bounds on this latter quantity. The proof of these theorems is presented in Section 6.

The first theorem bounds the expected value of $\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$ when $\mathcal{C}_\Upsilon$ is viewed as a random variable. Interestingly, this bound is essentially the same as the classical bound for the i.i.d. setting [13], except that the value of n from the classical bound is replaced by $\lfloor n/k \rfloor$, which is perhaps not surprising as it is the maximum number of non-overlapping size-k subsets of C_Υ.

Theorem 1. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a global example and $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let us define $\Upsilon = \aleph\langle\mathcal{C}_\Upsilon\rangle$. Then for any hypothesis class $\mathcal{H}$ of functions $f : \Omega(\mathcal{L}, k) \to \{0,1\}$ with finite VC-dimension d, the following holds:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|\right] \leq 2 \cdot \sqrt{\frac{2d \log\left(2e\lfloor n/k\rfloor/d\right)}{\lfloor n/k \rfloor}}$$

The second theorem provides a tail bound for $P\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)| \geq \varepsilon\right]$. We note that the bound on expected error from Theorem 1 cannot be derived from Theorem 2, although a different bound on expected error with looser constants could be derived from Theorem 2.

Theorem 2. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a global example and $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let us define $\Upsilon = \aleph\langle\mathcal{C}_\Upsilon\rangle$. Then for any hypothesis class $\mathcal{H}$ of functions $f : \Omega(\mathcal{L}, k) \to \{0,1\}$ with finite VC-dimension d, the following holds for any $0 < \varepsilon \leq 1$:*

$$P\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)| \geq \varepsilon\right]$$
$$\leq \exp\left(-\frac{\lfloor n/k\rfloor\varepsilon^2}{4}\right) + \varepsilon\sqrt{8\pi\lfloor n/k\rfloor}\left(\frac{2e\lfloor n/k\rfloor}{d}\right)^d \cdot \exp\left(-\frac{\lfloor n/k\rfloor\varepsilon^2}{8}\right)$$

Up to somewhat looser constants, the tail bound from Theorem 2 can be shown to also have the same form as the existing VC tail bounds [16]. In particular, the bound implies the following simpler, albeit looser bound:

$$P\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)| \geq \varepsilon\right]$$

$$\leq \left(1 + \sqrt{8\pi \lfloor n/k \rfloor} \left(\frac{2e\lfloor n/k \rfloor}{d}\right)^d\right) \cdot \exp\left(-\frac{\lfloor n/k \rfloor \varepsilon^2}{8}\right).$$

5 Related Work

There have been several works studying theoretical properties of various statistical relational learning settings. Dhurandhar and Dobra [3] derived Hoeffding-type inequalities for classifiers trained with relational data. However, there are several important differences with our work. First, their bounds are not VC-type bounds. Moreover, their results, based on restricting the independent interactions of data points, cannot be applied in our setting, which is more general than the one they consider. Certain other statistical properties of learning have also been studied for SRL models. For instance, Xiang and Neville [19] studied consistency of estimation in a certain relational learning setting.

From a different perspective, abstracting from the relational logic setting, our results can also be seen as bounds for uniform deviations of U-statistics [4] under sampling *without* replacement. Not many results are known for this particular setting in the literature. One exception is the work of Nandi and Sen [11] who only derived bounds on variance in this setting. It is not possible to derive our results from theirs. In particular, we need Chernoff-type bounds whereas the variance bounds from their work would only give us Chebyshev-type bounds. A more thoroughly studied setting is the estimation of U-statistics under sampling *with* replacement. Clémencon, Lugosi and Vayatis [1] derived among others[6] VC-inequalities in a setting similar to ours, but under sampling with replacement, which makes their analysis simpler. However, such an assumption would not make sense in the relational learning setting where it would mean, for instance, that we would end up with multiple copies of the same individual (e.g. ending up with social networks in which the same person can occur multiple times).

6 Derivation of the Bounds

In this section, we prove Theorems 1 and 2 using a series of lemmas. First, in Section 6.1, we define a sampling process for generating vectors containing $\lfloor n/k \rfloor$ size-k fragments of Υ. The sampling process has two important properties. First, the fragments in each of the vectors are distributed as size-k fragments sampled i.i.d. from $\aleph$ (assuming Υ is sampled as in Definition 1). Second, the average of the estimates of $Q_{\aleph,k}(f)$ computed from the vectors converges to $Q_{\Upsilon,k}(f)$. These two properties allow us to use the sampling process to derive a bound on expected value of the random variable $\sup_{f \in \mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$ in Section 6.2, which finishes the proof of Theorem 1.

The proof of Theorem 2 is a bit more involved. First, in Section 6.3, we derive bounds on the moment-generating function of a random variable that can be obtained if we only know its tail bounds. Then, in Section 6.4, we combine the results from the preceding sections to prove Theorem 2. In particular, we use the bound moment-generating function

[6] The main results of [1] are bounds that assume a certain ’low-noise’ condition. Although they only derived bounds for the case $k = 2$ (in our notation), the results directly related to ours can be extended for larger k’s as well.

to obtain a tail bound on the estimates of $Q_{\aleph,k}(f)$ by exploiting a trick that is sometimes called *average of sums-of-i.i.d blocks* [1].

6.1 Extracting Independent Samples

In this section we describe a sampling process that allows us to obtain $\lfloor n/k \rfloor$ samples from $\varUpsilon$ that are distributed as i.i.d. samples from $\aleph$, assuming $\varUpsilon$ is sampled as in Definition 1.

Lemma 1. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a global example. Let $0 \leq n \leq |\mathcal{C}_\aleph|$, $q \geq 1$ and $1 \leq k \leq n$ 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 $\mathcal{I}' = \{1, 2, \ldots, |\mathcal{C}_\aleph|\}$ and let $\mathbf{Y}_j = (\mathcal{S}'_{j,1}, \mathcal{S}'_{j,2}, \ldots, \mathcal{S}'_{j,\lfloor \frac{n}{k} \rfloor})$, for $1 \leq j \leq q$, be vectors sampled by the following process:*

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

Then the following holds:

1. *The random vectors $\mathbf{X}$ and $\mathbf{Y}_j$ have the same distribution for any $1 \leq j \leq q$.*
2. *For any function $f : \varOmega(\mathcal{L}, k) \rightarrow [0,1]$ it holds:*

$$P\left[\left|Q_{\varUpsilon,k}(f) - \frac{1}{q\lfloor n/k \rfloor}\sum_{j=1}^{q}\sum_{i=1}^{\lfloor n/k \rfloor} f\left(\varUpsilon\langle \mathcal{S}'_{j,i}\rangle\right)\right| \geq \epsilon\right] \leq 2\exp\left(-2q\varepsilon^2\right)$$

Proof. The first part of the proof follows immediatelly from Lemma 3 in [8] (which, for completeness, we reprove in the online[7] appendix as Lemma 5). For the second part, we may first notice that, after $\mathcal{C}_\varUpsilon$ is sampled and fixed, $Q_{\varUpsilon,k}(f) = \mathbb{E}\left[f\left(\varUpsilon\langle \mathcal{S}'_{j,i}\rangle\right)\right]$, as the probability of $\mathcal{S}'_{j,i}$ being a particular size-k subset of $\mathcal{C}_\varUpsilon$ is the same for all such subsets. The second part can then be shown by applying Hoeffding inequality to q i.i.d. samples $\frac{1}{\lfloor n/k \rfloor}\sum_{i=1}^{\lfloor n/k \rfloor} f(\varUpsilon\langle \mathcal{S}_{j,i}\rangle)$, $j = 1, 2, \ldots, q$, which have the same expected value $Q_{\varUpsilon,k}(f)$. □

At this point, one might wonder if the above lemma already gives us a way to find VC-type bounds for relational data, based on the following strategy: sample $\lfloor n/k \rfloor$ size-k fragments from a given training relational example $\varUpsilon$ using the procedure defined in Lemma 1 and use this set of fragments as our training data. Although this would allow us to use standard bounds that are known for learning from i.i.d. data [15], there are two problems with this approach. The first problem is that in reality we do not always

[7] `https://arxiv.org/abs/1804.06188`

know the size of the global example $\aleph$ and hence we do not know how to get a sample of $\lfloor n/k \rfloor$ size-k sets that behaves as an independent sample from $\aleph$ (noting that we need to know the size of $\aleph$ to define the set $\mathcal{I}'$ in Lemma 1). The second problem is that there are cases where only sampling the $\lfloor n/k \rfloor$ samples is sub-optimal from the point of view of statistical power, as we illustrate in the next example.

Example 4. Consider a global structure which takes the form of a large directed graph, and assume that we are interested in estimating the probability that the formula $\exists X, Y : \mathit{edge}(X, Y)$ holds for a fragment of the structure induced by two randomly sampled nodes. Assume furthermore that the given graph was generated by sampling (directed) edges independently with some probability p. The probability that $\exists X, Y : \mathit{edge}(X, Y)$ holds for any two nodes will thus correspond to some value p^* close to $1 - (1-p)^2$. As we will see, given a training fragment induced by n nodes from this graph, we can only generate $\lfloor \frac{n}{2} \rfloor$ samples that behave like i.i.d. samples. In this case, a more accurate estimate of p^* can be obtained by using all size-2 fragments of the training fragment.

Nonetheless, the strategy based on sampling $\lfloor n/k \rfloor$ size-k fragments may actually be optimal in the worst case as we illustrate in the next example.

Example 5. Let us again consider the setting from Example 4, which we can now describe more formally. In particular, assume that $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ represents a large directed graph. Let $k = 2$ and $\Phi = \{\exists X, Y : \mathit{edge}(X, Y)\}$. Let Υ be a relational example sampled uniformly from $\aleph$ (i.e. $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$ where $\mathcal{C}_\Upsilon$ is sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$). Let us now, in contrast to the assumption underlying Example 4, assume that the directed graph was constructed using the following process. For all nodes v, we flip a biased coin with probability of heads being q. If it lands heads, we add a directed edge from v to all other nodes. In this case[8], $Q_{\aleph,k}(\Phi) = p' \approx 1 - (1-q)^2$. The main difference with the setting from Example 4 is that estimating p' now effectively corresponds to estimation of a property of nodes, as we are also able to recover p' by observing how many nodes have at least one outgoing edge. However, this also means that the effective sample size in this case only grows linearly with the number of vertices (as opposed to quadratically in Example 4). This, at least asymptotically (up to a multiplicative constant), is a worst-case scenario as the number of independent samples that we are able to obtain using Lemma 1 also grows linearly with the number of vertices in the sample Υ (i.e. linearly with $|\mathcal{C}_\Upsilon|$).

6.2 Bounding Expected Error

In this section we use the results from Section 6.1 to obtain a bound on the expected value of $\sup_{f \in H} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|$.

Lemma 2. *Let $\aleph = (\mathcal{A}_\aleph, \mathcal{C}_\aleph)$ be a global example and $\mathcal{C}_\Upsilon$ be sampled uniformly from all size-n subsets of $\mathcal{C}_\aleph$ and let us define $\Upsilon = \aleph\langle \mathcal{C}_\Upsilon \rangle$. Let $\mathbf{Y}_j = (\mathcal{S}'_{j,1}, \dots, \mathcal{S}'_{j,\lfloor \frac{n}{k} \rfloor})$,*

[8] More formally, the following holds, assuming $\aleph$ is generated by the respective random processes. In the setting from Example 4 we have $\mathbb{E}_\aleph[Q_{\aleph,k}(\Phi)] = 1 - (1-p)^2$ and in the setting from this example we have $\mathbb{E}_\aleph[Q_{\aleph,k}(\Phi)] = 1 - (1-q)^2$.

where $1 \leq j \leq q$, be random vectors sampled as in Lemma 1. Then for any hypothesis class $\mathcal{H}$ of functions $f : \Omega(\mathcal{L}, k) \to \{0, 1\}$ with finite VC-dimension d, the following holds:

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|\right] \leq \lim_{q \to \infty} \mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| Q_{\aleph,k}(f) - \frac{1}{q \cdot \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right|\right]$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f \in \mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)|\right] \\
&\quad = \lim_{q \to \infty} \mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| Q_{\aleph,k}(f) - \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) \right.\right. \\
&\qquad\qquad \left.\left. + \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) - Q_{\Upsilon,k}(f) \right|\right] \\
&\quad \leq \lim_{q \to \infty} \mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| Q_{\aleph,k}(f) - \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) \right|\right] \\
&\qquad\qquad + \lim_{q \to \infty} \mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) - Q_{\Upsilon,k}(f) \right|\right] \qquad (1)
\end{aligned}$$

To finish the proof, we show that the last summand in (1) is zero. To this end, first note that it follows from Remark 1 that the supremum only needs to be taken over a finite number t of hypotheses, one from each equivalence class of functions that are equal on all size-k subsets of $\mathcal{C}_\aleph$. Together with Lemma 1 and the union bound on the finitely many equivalence classes, we find

$$P\left[\sup_{f \in \mathcal{H}} \left| \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) - Q_{\Upsilon,k}(f) \right| \geq \varepsilon \right] \leq 2 \cdot t \cdot \exp\left(-2q\varepsilon^2\right)$$

Then it follows using $\mathbb{E}[X] = \int_0^1 P[X \geq x]dx$ (assuming $P[X \in [0;1]] = 1$) that

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| \left(\frac{1}{q \left\lfloor \frac{n}{k} \right\rfloor} \sum_{i=1}^{q} \sum_{\mathcal{S} \in \mathbf{Y}_i} f(\Upsilon\langle \mathcal{S} \rangle) \right) - Q_{\Upsilon,k}(f) \right|\right] \leq \int_0^1 2 \cdot t \cdot \exp\left(-2qx^2\right) dx.$$

Finally, noticing that $\lim_{q \to \infty} \int_0^1 2 \cdot t \cdot \exp\left(-2qx^2\right) dx = 0$ finishes the proof. □

Lemma 3. *Suppose $\mathbf{Y}_j = (\mathcal{S}'_{j,1}, \dots, \mathcal{S}'_{j,\lfloor \frac{n}{k} \rfloor})$ is a random vector sampled as in Lemma 1. Then for any hypothesis class of functions $f : \Omega(\mathcal{L}, k) \to \{0, 1\}$ with VC-dimension d we have:*

$$P\left[\sup_{f \in \mathcal{H}} \left| Q_{\aleph,k}(f) - \frac{1}{\lfloor n/k \rfloor} \sum_{\mathcal{S} \in \mathbf{Y}_j} f(\Upsilon\langle \mathcal{S} \rangle) \right| \geq \varepsilon \right] \leq 4 \left(\frac{2e\lfloor n/k \rfloor}{d} \right)^d \exp\left(-\frac{\lfloor n/k \rfloor \varepsilon^2}{8} \right)$$

and

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}}\left|Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right]\leq 2\sqrt{\frac{2d\log\left(2e\lfloor n/k\rfloor/d\right)}{\lfloor n/k\rfloor}}$$

Proof. Since $\mathcal{S}'_{j,1},\dots,\mathcal{S}'_{j,\lfloor\frac{n}{k}\rfloor}$ are sampled in an i.i.d. way, the classical VC inequality applies [16]. The expected value bound can be derived from the bound (6.4) in [13][9]. □

We are now ready to prove Theorem 1.

Proof (of Theorem 1). Let $\mathbf{Y}_j=(\mathcal{S}'_{j,1},\dots,\mathcal{S}'_{j,\lfloor\frac{n}{k}\rfloor})$, where $1\leq j\leq q$ for a given integer q, be random vectors sampled as in Lemma 1. First, using Lemma 2 for the first step, we find

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in H}\left|Q_{\aleph,k}(f)-Q_{\Upsilon,k}(f)\right|\right]\\
&\quad\leq\lim_{q\to\infty}\mathbb{E}\left[\sup_{f\in H}\left|Q_{\aleph,k}(f)-\frac{1}{q}\sum_{j=1}^{q}\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right]\\
&\quad=\lim_{q\to\infty}\mathbb{E}\left[\sup_{f\in H}\left|\frac{1}{q}\sum_{j=1}^{q}\left(Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right)\right|\right]\\
&\quad\leq\lim_{q\to\infty}\mathbb{E}\left[\frac{1}{q}\sup_{f\in H}\sum_{j=1}^{q}\left|Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right]\\
&\quad\leq\lim_{q\to\infty}\mathbb{E}\left[\frac{1}{q}\sum_{j=1}^{q}\sup_{f\in H}\left|Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right]\\
&\quad=\lim_{q\to\infty}\frac{1}{q}\sum_{j=1}^{q}\mathbb{E}\left[\sup_{f\in H}\left|Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right]\\
&\quad=\mathbb{E}\left[\sup_{f\in H}\left|Q_{\aleph,k}(f)-\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_1}f(\Upsilon\langle\mathcal{S}\rangle)\right|\right] \qquad (2)
\end{aligned}$$

Note that the last equality is a consequence of Lemma 1, from which it among others follows that all $\mathbf{Y}_j$'s have the same distribution. In other words, all the q expected values are equal. Finally, we can use Lemma 3 to bound (2) which finishes the proof. □

It is also possible to get rid of the logarithmic factor in the bound on expected error. However, as mentioned in [2], such bounds are worse up to very large training set sizes due to the increased constant factors.

[9] The specific form that we use here can be found in the lecture notes of Philippe Rigollet `https://bit.ly/2H89wPn`.

6.3 From Tail Bounds to Moment-Generating Functions

In this section, we derive bounds on the moment-generating function of a random variable from its tail bounds.

Lemma 4. *For a non-negative random variable X, if there exist constants $C \geq e$ and $B > 0$ such that*

$$P[X \geq t] \leq C \exp(-t^2/B) \qquad \forall t \geq 0,$$

then for any $\lambda > 0$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq 1 + \lambda C \sqrt{\pi B} \exp\left(\frac{\lambda^2 B}{4}\right)$$

Proof. We have:

$$\begin{aligned}\mathbb{E}\left[X^p\right] &= \int_0^\infty \mathrm{P}\left(X^p \geq u\right) du = \int_0^\infty \mathrm{P}\left(X^p \geq t^p\right) \cdot p \cdot t^{p-1} dt \\ &= \int_0^\infty \mathrm{P}\left(X \geq t\right) \cdot p \cdot t^{p-1} dt \leq \int_0^\infty C \cdot e^{-t^2/B} \cdot p \cdot t^{p-1} dt\end{aligned}$$

Next, for the moment-generating function, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X\right)\right] &\leq 1 + \sum_{p=1}^{\infty} \frac{\lambda^p \mathbb{E}\left[X^p\right]}{p!} \leq 1 + \sum_{p=1}^{\infty} \frac{\lambda^p \int_0^\infty C \cdot e^{-t^2/B} \cdot p \cdot t^{p-1} dt}{p!} \\ &\leq 1 + C \int_0^\infty e^{-t^2/B} \cdot \sum_{p=1}^{\infty} \frac{\lambda^p p \cdot t^{p-1}}{p!} dt \\ &= 1 + C\lambda \int_0^\infty e^{-t^2/B} \cdot \sum_{p=0}^{\infty} \frac{\lambda^p \cdot t^p}{p!} dt \\ &= 1 + C\lambda \int_0^\infty e^{-t^2/B} \cdot e^{t\lambda} dt = 1 + C\lambda \int_0^\infty e^{-\frac{\left(t - \frac{1}{2}\lambda B\right)^2}{B} + \frac{\lambda^2 B^2}{4}} dt \\ &= 1 + C\lambda e^{\frac{\lambda^2 B^2}{4}} \int_0^\infty e^{-\frac{\left(t - \frac{1}{2}\lambda B\right)^2}{B}} dt \\ &= 1 + \frac{1}{2} C\lambda \sqrt{\pi B} e^{\frac{\lambda^2 B^2}{4}} \left(\operatorname{erf}\left(\frac{\lambda\sqrt{B}}{2}\right) + 1\right) \\ &\leq 1 + C\lambda\sqrt{\pi B} \exp\left(\frac{\lambda^2 B}{4}\right)\end{aligned}$$

Note that it is easy to check that all the series in the above derivation converge absolutely. The Fubini-Tonelli theorem justifies the change of order of summation and integration. □

6.4 From Moment-Generating Functions to Tail Bounds

We can now finish the proof of our main result, Theorem 2.

Proof (of Theorem 2). Let $\mathbf{Y}_j = (\mathcal{S}'_{j,1}, \dots, \mathcal{S}'_{j,\lfloor \frac{n}{k} \rfloor})$, for $1 \leq j \leq q$, be random vectors sampled as in Lemma 1. For convenience, let us also define

$$R^{(q)}_{\Upsilon}(f) = \frac{1}{q} \sum_{j=1}^{q} \frac{1}{\lfloor n/k \rfloor} \sum_{\mathcal{S} \in \mathbf{Y}_j} f(\Upsilon\langle \mathcal{S} \rangle).$$

First, we have

$$\begin{aligned}
&P\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)| \geq \varepsilon\right] \\
&\qquad = P\left[\sup_{f\in\mathcal{H}} \left\{\left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f) + R^{(q)}_{\Upsilon}(f) - Q_{\Upsilon,k}(f)\right|\right\} \geq \varepsilon\right] \\
&\qquad \leq P\left[\sup_{f\in\mathcal{H}} \left\{\left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right| + \left|R^{(q)}_{\Upsilon}(f) - Q_{\Upsilon,k}(f)\right|\right\} \geq \varepsilon\right] \\
&\qquad \leq P\left[\sup_{f\in\mathcal{H}} \left\{\left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right|\right\} + \sup_{f\in\mathcal{H}} \left\{\left|R^{(q)}_{\Upsilon}(f) - Q_{\Upsilon,k}(f)\right|\right\} \geq \varepsilon\right]
\end{aligned}$$

It follows from the fact that the supremum needs to be taken only over the finitely many equivalence classed of $\mathcal{H}$ on $\aleph$ and from Lemma 1 (see the discussion in the proof of Lemma 5) that for any $\varepsilon^* > 0$ and $\delta^* > 0$ there is an integer q_0 such that for all $q \geq q_0$:

$$P\left[\sup_{f\in\mathcal{H}} \left\{\left|R^{(q)}_{\Upsilon}(f) - Q_{\Upsilon,k}(f)\right|\right\} \geq \varepsilon^*\right] \leq \delta^*.$$

Hence, for any $\varepsilon^* > 0$, $\delta^* > 0$ and a suitably large $q \geq q_0$ we have

$$\begin{aligned}
&P\left[\sup_{f\in\mathcal{H}} \left\{\left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right|\right\} + \sup_{f\in\mathcal{H}} \left\{\left|R^{(q)}_{\Upsilon}(f) - Q_{\Upsilon,k}(f)\right|\right\} \geq \varepsilon\right] \\
&\qquad\qquad \leq P\left[\sup_{f\in\mathcal{H}} \left\{\left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right|\right\} \geq \varepsilon - \varepsilon^*\right] + \delta^*.
\end{aligned}$$

Taking the limit $q_0 \to \infty$ we obtain

$$P\left[\sup_{f\in\mathcal{H}} |Q_{\aleph,k}(f) - Q_{\Upsilon,k}(f)| \geq \varepsilon\right] \;\leq\; \lim_{q\to\infty} P\left[\sup_{f\in\mathcal{H}} \left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right| \geq \varepsilon\right]$$

Next we need to bound the right-hand side of the above inequality. For any q we have

$$P\left[\sup_{f\in\mathcal{H}} \left|Q_{\aleph,k}(f) - R^{(q)}_{\Upsilon}(f)\right| \geq \varepsilon\right]$$

$$
\begin{aligned}
&= P\left[\sup_{f\in\mathcal{H}}\left|Q_{\aleph,k}(f) - \frac{1}{q}\sum_{j=1}^{q}\frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j} f(\varUpsilon\langle\mathcal{S}\rangle)\right| \geq \varepsilon\right] \\
&= P\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{q}\sum_{j=1}^{q}\left(Q_{\aleph,k}(f) - \frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j} f(\varUpsilon\langle\mathcal{S}\rangle)\right)\right| \geq \varepsilon\right] \\
&\leq P\left[\sup_{f\in\mathcal{H}}\frac{1}{q}\sum_{j=1}^{q}\left|Q_{\aleph,k}(f) - \frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j} f(\varUpsilon\langle\mathcal{S}\rangle)\right| \geq \varepsilon\right] \\
&\leq P\left[\frac{1}{q}\sum_{j=1}^{q}\sup_{f\in\mathcal{H}}\left|Q_{\aleph,k}(f) - \frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j} f(\varUpsilon\langle\mathcal{S}\rangle)\right| \geq \varepsilon\right]
\end{aligned}
$$

Let us denote

$$T_j = \sup_{f\in\mathcal{H}}\left|Q_{\aleph,k}(f) - \frac{1}{\lfloor n/k\rfloor}\sum_{\mathcal{S}\in\mathbf{Y}_j} f(\varUpsilon\langle\mathcal{S}\rangle)\right|.$$

Combining Lemma 3 and Lemma 4, we can bound $\mathbb{E}\left[\exp\left(\lambda T_j\right)\right]$ as

$$\mathbb{E}\left[\exp\left(\lambda T_j\right)\right] \leq 1 + 4\lambda\sqrt{\frac{8\pi}{\lfloor n/k\rfloor}}\left(\frac{2e\lfloor n/k\rfloor}{d}\right)^d \exp\left(\frac{2\lambda^2}{\lfloor n/k\rfloor}\right).$$

Let us denote $T = \frac{1}{q}\sum_{j=1}^{q} T_j$. We use the observation from [5] that due to Jensen's inequality and linearity of expectation

$$\mathbb{E}\left[\exp\left(\lambda T\right)\right] \leq \frac{1}{q}\sum_{j=1}^{q}\mathbb{E}\left[\exp\left(\lambda T_j\right)\right] = \mathbb{E}\left[\exp\left(\lambda T_1\right)\right].$$

Next we obtain a bound on $P[T \geq \varepsilon]$ from the bound on $\mathbb{E}\left[\exp\left(\lambda T\right)\right]$. In particular, for positive λ, we have

$$
\begin{aligned}
P[T \geq \varepsilon] &= P[e^{\lambda\cdot X} \geq e^{\lambda\cdot\varepsilon}] \leq e^{-\lambda\cdot\varepsilon}\mathbb{E}\left[e^{\lambda\cdot T}\right] \\
&\leq e^{-\lambda\cdot\varepsilon}\left(1 + 4\lambda\sqrt{\frac{8\pi}{\lfloor n/k\rfloor}}\left(\frac{2e\lfloor n/k\rfloor}{d}\right)^d \exp\left(\frac{2\lambda^2}{\lfloor n/k\rfloor}\right)\right).
\end{aligned}
$$

where the Markov inequality was used for the third step. Since the above bound holds for any q, it also holds in the limit. Next, we can plug in $\lambda := \frac{\varepsilon\cdot\lfloor n/k\rfloor}{4}$ and obtain:

$$P[T \geq \varepsilon] \leq \exp\left(-\frac{\lfloor n/k\rfloor\varepsilon^2}{4}\right) + \varepsilon\sqrt{8\pi\lfloor n/k\rfloor}\left(\frac{2e\lfloor n/k\rfloor}{d}\right)^d \cdot \exp\left(-\frac{\lfloor n/k\rfloor\varepsilon^2}{8}\right).$$

□

7 Concluding Remarks

We have derived VC-dimension based bounds which can be applied in relational learning settings where one may assume that the training data (i.e. some given relational structure) was obtained from a larger relational structure by sampling without replacement. This includes many of the typical application settings in which, for instance, Markov logic networks are used. The considered bounds are useful, among others, for the analysis of relational marginal problems [8] and PAC-reasoning in relational domains [7].

There are several interesting avenues for future work. First, in this paper, we have not studied the realizable learning case for which, at least in the classical i.i.d. case, one can obtain faster convergence rates. It would be interesting to extend our results into the realizable case. Similarly, it would be of interest to study bounds under low-noise conditions [1], which sit somewhere between the realizable case and the case studied in this paper. Another natural direction for future work would be to extend the PAC-Bayesian setting into relational learning, as the bounds that are derived in this setting tend to be tighter in practice [10].

References

1. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and empirical minimization of u-statistics. The Annals of Statistics pp. 844–874 (2008)
2. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition, Stochastic Modelling and Applied Probability, vol. 31. Springer (1996)
3. Dhurandhar, A., Dobra, A.: Distribution-free bounds for relational classification. Knowledge and information systems **31**(1), 55–78 (2012)
4. Hoeffding, W.: A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics pp. 293–325 (1948)
5. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American statistical association **58**(301), 13–30 (1963)
6. Juba, B.: Implicit learning of common sense for reasoning. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. pp. 939–946 (2013)
7. Kuželka, O., Wang, Y., Davis, J., Schockaert, S.: PAC-reasoning in relational domains. In: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, UAI 2018 (2018)
8. Kuželka, O., Wang, Y., Davis, J., Schockaert, S.: Relational marginal problems: Theory and estimation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) (2018)
9. Kuželka, O., Davis, J., Schockaert, S.: Induction of interpretable possibilistic logic theories from relational data. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 1153–1159 (2017)
10. Langford, J., Shawe-Taylor, J.: PAC-Bayes & margins. In: Proceedings of the Annual Conference on Neural Information Processing Systems. pp. 423–430 (2002)
11. Nandi, H., Sen, P.: 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 (1963)
12. Rocktäschel, T., Riedel, S.: End-to-end differentiable proving. In: Proceedings of the Annual Conference on Neural Information Processing Systems. pp. 3791–3803 (2017)
13. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)

14. Valiant, L.G.: Knowledge infusion. In: Proceedings of the 21st National Conference on Artificial Intelligence. pp. 1546–1551 (2006)
15. Vapnik, V.: The Nature of Statistical Learning Theory. Springer: New York (2000)
16. Vapnik, V., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16**(2), 264 (1971)
17. Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R., Borgwardt, K.M.: Graph kernels. Journal of Machine Learning Research **11**, 1201–1242 (2010)
18. Šourek, G., Aschenbrenner, V., Železný, F., Kuželka, O.: Lifted relational neural networks. In: Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches. (2015)
19. Xiang, R., Neville, J.: Relational learning with one network: An asymptotic analysis. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. pp. 779–788 (2011)