Implicitly Learning to Reason in First-Order Logic

Implicitly Learning to Reason in First-Order Logic

Vaishak Belle[1], Brendan Juba[2]

[1]University of Edinburgh, UK & Alan Turing Institute, UK
vaishak@ed.ac.uk
[2]Department of Computer Science & Engineering, Washington University in St. Louis, USA
bjuba@wustl.edu

Abstract

We consider the problem of answering queries about formulas of first-order logic based on background knowledge partially represented explicitly as other formulas, and partially represented as examples independently drawn from a fixed probability distribution. PAC semantics, introduced by Valiant, is one rigorous, general proposal for learning to reason in formal languages: although weaker than classical entailment, it allows for a powerful model theoretic framework for answering queries while requiring minimal assumptions about the form of the distribution in question. To date, however, the most significant limitation of that approach, and more generally most machine learning approaches with robustness guarantees, is that the logical language is ultimately essentially propositional, with finitely many atoms. Indeed, the theoretical findings on the learning of relational theories in such generality have been resoundingly negative. This is despite the fact that first-order logic is widely argued to be most appropriate for representing human knowledge. In this work, we present a new theoretical approach to robustly learning to reason in first-order logic, and consider universally quantified clauses over a countably infinite domain. Our results exploit symmetries exhibited by constants in the language, and generalize the notion of implicit learnability to show how queries can be computed against (implicitly) learned first-order background knowledge.

1 Introduction

The tension between *deduction* and *induction* is perhaps the most fundamental issue in areas such as philosophy, cognition and artificial intelligence. The deduction camp concerns itself with questions about the expressiveness of formal languages for capturing knowledge about the world, together with proof systems for reasoning from such *knowledge bases*. The learning camp attempts to generalize from examples about partial descriptions about the world. In an influential paper, [Valiant, 2000] recognized that the challenge of learning should be integrated with deduction. In particular, he proposed a semantics to capture the quality possessed by the output of (probably approximately correct) PAC-learning algorithms when formulated in a logic. Although weaker than classical entailment, it allows for a powerful model theoretic framework for answering queries.

From the standpoint of learning an expressive logical knowledge base and reasoning with it, most PAC results are somewhat discouraging. For example, in agnostic learning [Kearns *et al.*, 1994] where one does not require examples (drawn from an arbitrary distribution) to be fully consistent with learned sentences, efficient algorithms for learning conjunctions would yield an efficient algorithm for PAC-learning DNF (also over arbitrary distributions), which current evidence suggests to be intractable [Daniely and Shalev-Shwartz, 2016]. Thus, it is not surprising that when it comes to first-order logic (FOL), very little work tackles the problem in a general manner. This is despite the fact that FOL is widely argued to be most appropriate for representing human knowledge (e.g., [McCarthy and Hayes, 1969; Moore, 1982; Levesque and Lakemeyer, 2001]). For example, [Cohen and Hirsh, 1994] consider the problem of the learnability of description logics with equality constraints. While description logics are already restricted fragments of FOL in only allowing unary and some binary predicates, it is shown that such a fragment cannot be tractably learned, leading to the identification of syntactic restrictions for learning from positive examples alone. Analogously, when it comes to the learning of logic programs [Cohen and Page, 1995], which in principle may admit infinitely many terms, syntactic restrictions are also typical [De Raedt and Džeroski, 1994].

In this work, we present new results on learning to reason in FOL knowledge bases. In particular, we consider the problem of answering queries about FOL formulas based on background knowledge partially represented explicitly as other formulas, and partially represented as examples independently drawn from a fixed probability distribution. Our results are based on a surprising observation made in [Juba, 2013] about the advantages of eschewing the explicit construction of a hypothesis, leading to a paradigm of *implicit learnability*. Not only does it enable a form of agnostic learning while circumventing known barriers, it also avoids the design of an often restrictive and artificial choice for representing hypotheses. (See, for example, [Khardon and Roth, 1999], which is similar in spirit in allowing declarative background knowledge but only permits constant-width clauses.) In particular, implicit learning allows such learning from partially observed examples, which is commonplace when knowledge bases and/or queries address entities and relations not observed in the data used for learning.

That work was limited to the propositional setting, however. Here, we develop a first-order logical generalization. Since reasoning in full FOL is undecidable we need to consider a fragment, but the fragment we identify and are able to learn and reason with is expressive and powerful. Consider that standard databases correspond to a maximally consistent and finite set of literals: every relevant atom is known to be true and stored in the database, or known to be false, inferred by (say) negation as failure. Our fragment corresponds to a consistent but infinite set of ground clauses, not necessarily maximal. To achieve the generalization, we revisit the PAC semantics and exploit symmetries exhibited by constants in the language. Moreover, the underlying language is general in the sense that no restrictions are posed on clause length, predicate arity, and other similar technical devices seen in PAC results. We hope the simplicity of the framework is appealing to the readers and hope our results will renew interest in learnability for expressive languages with quantificational power.

We remark that our sole focus is in PAC-semantics approaches, but there are also other families of methods for unifying statistical and logical representations, that fall under the banner of *statistical relational learning* (SRL) (e.g., [Kersting *et al.*, 2011]). SRL includes widely used formalisms such as Markov Logic Networks [Richardson and Domingos, 2006] and frameworks such as Inductive Logic Programming [Muggleton and De Raedt, 1994]. Generally speaking, there are significant differences to PAC-semantics approaches, such as in terms of the learning regime, the notion of correctness and the underlying algorithmic machinery. For example, Markov Logic Networks use approximate maximum-likelihood learning strategies to capture the distribution of the data, whereas in PAC formulations, one considers an arbitrary unknown distribution over the data and studies the question of what formulas are learnable whilst costing for the number of examples needed to be sampled from that distribution. Of course, there is much to be gained by attempting to integrate these communities; see, for example, [Cohen and Page, 1995]. These differences notwithstanding, the learning of logical theories is usually restricted to finite-domain first-order logic, and so it is essentially propositional, and in that regard, our setting is significantly more challenging.

2 Logical Framework

Language: We let $\mathcal{L}$ be a first-order language with equality and relational symbols $\{P(x), \ldots, Q(x_1, \ldots, x_k), \ldots\}$, variables $\{x, y, z, \ldots\}$, and a countably infinite set of *rigid designators* or *names*, say, the set of natural numbers $\mathbb{N}$, serving as the domain of discourse for quantification. Well-defined formulas are constructed using logical connectives $\{\neg, \vee, \forall, \wedge, \exists, \supset\}$, as usual. Together with equality, names essentially realize an infinitary version of the unique-name assumption.[1]

The set of (ground) atoms is obtained as:[2] ATOMS $= \{P(a_1, \ldots, a_k) \mid P \text{ is a predicate}, a_i \in \mathbb{N}\}$. We sometimes refer to elements of ATOMS as propositions, and ground formulas as propositional formulas. We will use p, q, e to denote atoms, and $\alpha, \beta, \phi, \psi$ to denote ground formulas.

Semantics: A $\mathcal{L}$-model M is a $\{0, 1\}$ assignment to the elements of ATOMS. Using $\models$ to denote satisfaction, the semantics for $\phi \in \mathcal{L}$ is defined as usual inductively, but with equality as identity: $M \models (a = b)$ iff a and b are the same names, and quantification understood substitutionally over all names in $\mathbb{N}$: $M \models \forall x \phi(x)$ iff $M \models \phi(a)$ for all $a \in \mathbb{N}$. We say that ϕ is *valid* iff for every $\mathcal{L}$-model M, $M \models \phi$. Let the set of all models be $\mathcal{M}$.

Representation: Like in standard FOL, reasoning over the full fragment of $\mathcal{L}$ is undecidable. Interestingly, owing to a fixed, albeit countably infinite, domain of discourse, the *compactness* property that holds for classical first-order logic *does not hold* in general [Levesque, 1998]. For example, $\{\exists x P(x), \neg P(1), \neg P(2), \ldots\}$ is an unsatisfiable theory for which every finite subset is indeed satisfiable. However, as identified in [Belle, 2017], and earlier in [Lakemeyer and Levesque, 2002], the case of disjunctive knowledge is more manageable. In particular, we will be interested in learning and reasoning with incomplete knowledge bases with disjunctive information [Belle, 2017]:

Definition 1: An *acceptable* equality is of the form $x = a$, where x is any variable and a any name. Let e range over formulas built from acceptable equalities and connectives $\{\neg, \vee, \wedge\}$. Let c range over quantifier-free disjunctions of (possibly non-ground) atoms. Let $\forall \phi$ mean the universal closure of ϕ. A formula of the form $\forall (e \supset c)$ is called a $\forall$-clause. A knowledge base (KB) Δ is proper$^+$ if it is a finite non-empty set of $\forall$-clauses. The *rank* of Δ is the maximum number of variables mentioned in any $\forall$-clause in Δ.

This fragment is very expressive. Consider that standard databases correspond to a maximally consistent and finite set of literals: every relevant atom is known to be true and stored in the database, or known to be false, inferred by (say) negation as failure. In contrast, such KBs correspond to a consistent but infinite set of ground clauses, not necessarily maximal.

Grounding: A ground theory is obtained from Δ by substituting variables with names. Suppose θ denotes a substitution. For any set of names $C \subseteq \mathbb{N}$, we write $\theta \in C$ to mean

[1] Our language $\mathcal{L}$ is essentially equivalent to standard FOL together a unique-name assumption for infinitely many constants [Levesque, 1998, Definition 3].

In general, the unique-name assumption does not rule out capturing uncertainty about the identity of objects; see [Giacomo *et al.*, 2011; Srivastava *et al.*, 2014], for example.

[2] Because equality is treated separately, atoms and clauses do not include equalities.

substitutions are only allowed wrt the names in C. Formally, we define:

- $\mathrm{GND}(\Delta) = \{c\theta \mid \forall(e \supset c) \in \Delta, \theta \in \mathbb{N}$ and $\models e\theta\}$;
- For $z \geq 0$, $\mathrm{GND}(\Delta, z) = \{c\theta \mid \forall(e \supset c) \in \Delta, \models e\theta, \theta \in Z\}$, where Z is the set of names mentioned in Δ plus z (arbitrary) new ones;
- For $C \subseteq \mathbb{N}$, $\mathrm{GND}(\Delta, C) = \{c\theta \mid \forall(e \supset c) \in \Delta, \models e\theta, \theta \in Z\}$ where Z is the set of names mentioned in Δ plus the names in C;
- $\mathrm{GND}^{-}(\Delta) = \mathrm{GND}(\Delta, z)$ where z is the rank of Δ.

Reasoning: Unfortunately, arbitrary reasoning with such KBs is also undecidable [Lakemeyer and Levesque, 2002, Theorem 7]. Various proposals have appeared to consider that problem: in [Lakemeyer and Levesque, 2002], for example, a sound but incomplete evaluation-based semantics is studied. In [Belle, 2017], it is instead shown that when the query is limited to ground formulas, we can reduce first-order entailment to propositional satisfiability:

Theorem 2: *[Belle, 2017] Suppose Δ is a proper$^+$ KB, and α is a ground formula. Then, $\Delta \models \alpha$ iff $\mathrm{GND}^{-}(\Delta \wedge \neg\alpha)$ is unsatisfiable.*

Here, the RHS of the iff is a propositional formula, obtained by a finite grounding, as defined above.

Example 3: Suppose $\Delta = \{\forall x(\mathit{Grad}(x) \vee \mathit{Prof}(x)), \forall x(x \neq \mathit{charles} \supset \mathit{Grad}(x))\}$ and the query is $\mathit{Grad}(\mathit{logan})$. Given that the KB's rank is 1, consider the grounding of the KB and the negated query wrt $\{\mathit{charles}, \mathit{logan}, \mathit{jean}\}$ (here *jean* is chosen arbitrarily). It is indeed unsatisfiable.

It is worth noting that the proof here (and in other proposals with $\mathcal{L}$-like languages [Levesque and Lakemeyer, 2001; Lakemeyer and Levesque, 2002; Liu and Levesque, 2005]) is established by setting up a bijection between names to show that all names other than those that appear in the finite grounding in the RHS behave "identically," and so for entailment purposes, it suffices to consider a finite set consisting of the constants already mentioned and a few extra ones. That idea can be traced back to [Levesque, 1998] (reformulated here for our purposes):

Theorem 4: *[Levesque, 1998] Suppose $\alpha = \forall x \phi(x)$ is a $\forall$-clause. (Its rank is 1.) Let C be the names mentioned in $\mathrm{GND}(\alpha, 1)$. Then for every $a \in \mathbb{N}$, there is a $b \in C$ such that $\models \phi(a)$ iff $\models \phi(b)$.*

The essence of Theorem 2 is to exploit this idea to show (reformulated here for our purposes):

Lemma 5: *[Belle, 2017] Suppose α is as above. If $\mathrm{GND}(\alpha, 1)$ is satisfiable, then so is $\mathrm{GND}(\alpha, z)$ for $z \geq 1$.*

Thus, we can extend a model that satisfies $\mathrm{GND}(\alpha, 1)$ to one that satisfies $\mathrm{GND}(\alpha)$, and so α itself. These observations will now lead to an appealing account for *implicit learnability* with proper$^+$ KBs.

3 Generalizing PAC-Semantics

Inductive generalization (as opposed to deduction) inherently has to cope with mistakes. Thus, the kind of knowledge produced by learning algorithms cannot hope to be valid in the traditional (Tarskian) sense, except in extreme cases, such as assuming we see every data point in a noise-free manner. The PAC semantics was introduced by Valiant [2000] to capture the quality possessed by the output of PAC-learning algorithms when formulated in a logic. In the classical propositional formulation, we suppose a propositional language with (say) n propositions, yielding a model theoretic space $\{0, 1\}^n$. We suppose that we observe examples independently drawn from a distribution D over $\{0, 1\}^n$. Then, suppose further that these examples enable a learning algorithm to find a formula ϕ. We cannot expect this formula to be valid in the traditional sense, as PAC-learning does not guarantee that the rule holds for every possible binding, only that ϕ so produced agrees with probability $1 - \epsilon$ wrt future examples drawn from the same distribution. This motivates a weaker notion of validity:

Definition 6: [$(1 - \epsilon)$-valid] Given a distribution D over $\{0, 1\}^n$, we say that a Boolean function F is $(1 - \epsilon)$-*valid* if $\Pr_{x \in D}[F(x) = 1] \geq 1 - \epsilon$. If $\epsilon = 0$, we say F is *perfectly valid*.

Thus far, the PAC semantics and its application to the formalization of robust logic-based learning has been limited to the propositional setting [Valiant, 2000; Michael, 2009; Juba, 2013], that is, where the learning vocabulary is finitely many atoms, and the background knowledge is essentially restricted to a propositional formula.[3] Generalizing that to the FOL case has to address, among other things, what $(1 - \epsilon)$-validity would like, how FOL formulas could be found by algorithms, and finally, how entailments can be computed. That is precisely our goal for this paper.

We start by proposing an extension of the PAC semantics for the infinitary structures constructed for $\mathcal{L}$, namely $\mathcal{M}$. For this, we will need to consider distributions on $\mathcal{M}$, which are defined as usual [Billingsley, 1995]: we take $\mathcal{M}$ to be the sample space (of elementary events), define a σ-algebra $\mathbf{M}$ to be a set of subsets of $\mathcal{M}$, which represent a collection of (not necessarily elementary) events, and a function $\Pr\colon \mathbf{M} \to [0, 1]$, which is the probability measure.

We are now ready to define $(1 - \epsilon)$-validity as needed in the PAC semantics.

Definition 7: Given a distribution Pr over $\mathcal{M}$, we say a formula $\phi \in \mathcal{L}$ is $(1 - \epsilon)$-valid iff $\Pr([\![\phi]\!]) \geq 1 - \epsilon$. If $\epsilon = 0$, then we say that ϕ is perfectly valid. Here, $[\![\phi]\!]$ for any closed formula $\phi \in \mathcal{L}$ denotes the set $\{M \in \mathcal{M} \mid M \models \phi\}$.

In practice, the most important use of the notion of validity is to check the entailment of a formula from a knowledge base, and by extension, the reader may wonder how that carries over from classical validity. As also observed in [Juba, 2013] (for the propositional case), the union bound allows classical reasoning to have a natural analogue in the PAC

[3] Valiant [2000] uses a fragment of FOL for which propositionalization is guaranteed to yield a small propositional formula, and only considers such a reduction to the propositional case.

semantics, shown below. Note that, as already mentioned, our assumption henceforth is that knowledge bases are proper$^+$, and queries are ground formulas, both in the context of reasoning as well as learning.

Proposition 8: *Let $\psi_1, \ldots, \psi_k$ be $\forall$-clauses such that each ψ_i is $(1-\epsilon_i)$-valid under a common distribution D for some $\epsilon_i \in [0,1]$. Suppose $\{\psi_1, \ldots, \psi_k\} \models \varphi$, for some ground formula φ. Then φ is $(1-\epsilon')$-valid under D for $\epsilon' = \sum_i \epsilon_i$.*

4 Partial Observability

The learning problem of interest here is to obtain knowledge about the distribution D, which, of course, is not revealed directly, but in the form of a set of examples. The examples in question are models independently drawn from D, and we are then interested in knowing whether a query α is $(1-\epsilon)$-valid. Intuitively, background knowledge Δ may be provided additionally and so the examples correspond to additional knowledge that the agent learns. This additional knowledge is never materialized in the form of $\mathcal{L}$-formulas, but is left implicit, as postulated first in [Juba, 2013].

When it comes to the examples themselves, however, we certainly cannot expect the examples to reveal the full nature of the world, and indeed, partial descriptions are commonplace in almost all applications [Michael, 2010]. In the case of $\mathcal{L}$, moreover, providing a full description may even be impossible in finite time. All of this motivates the following:

Definition 9: A partial model N maps ATOMS to $\{1, 0, *\}$. We say N is consistent with a $\mathcal{L}$-model M iff for all $p \in$ ATOMS, if $N[p] \neq *$ then $N[p] = M[p]$. Let $\mathcal{N}$ be the set of all partial models.

Essentially, our knowledge of D will be obtained from a set of partial models that are the examples.

Definition 10: A mask is a function θ that maps $\mathcal{L}$-models to partial models, with the property that for any $M \in \mathcal{M}$, $\theta(M)$ is consistent with M. A *masking process* Θ is a mask-valued random variable (i.e., a random function). We denote the distribution over partial models obtained by applying a masking process Θ to a distribution D over $\mathcal{L}$-models by $\Theta(D)$.

The definition of masking processes allows the hiding of entries to depend on the underlying example from D. Moreover, as discussed in [Juba, 2013] (for the propositional case), reasoning in PAC-Semantics from complete examples is trivial, whereas the hiding of all entries by a masking process means that the problem reduces to classical entailment. So, we expect examples to be of a sort that is in between these extremes. In particular, for the sake of tractable learning, we must consider formulas that can be evaluated efficiently from the partial models with high probability. This leads to a notion of *witnessing*.

Definition 11: We define a propositional formula $\phi \in \mathcal{L}$ to be witnessed to evaluate to true or false in a partial assignment N by induction as follows:

- an atom $Q(\vec{c})$ is witnessed to be true/false iff it is true/false respectively in N;
- $\neg\phi$ is witnessed true/false iff ϕ is witnessed false/true respectively;
- $\phi \vee \psi$ is witnessed true iff either ϕ or ψ is, and it is witnessed false iff both ϕ and ψ are witnessed false;
- $\phi \wedge \psi$ is witnessed true iff both ϕ and ψ are witnessed true, and it is witnessed false iff either ϕ or ψ is witnessed false;
- $\phi \supset \psi$ is witnessed true iff either ϕ is witnessed false or ψ is witnessed true, and it is witnessed false iff both ϕ is witnessed true and ψ is witnessed false.

We define a $\forall$-clause $\forall\vec{x}\phi(\vec{x})$ to be witnessed true in a partial model N for the set of names C if for every binding of $\vec{x}$ to names $\vec{c} \in C$, the resulting ground clause $\phi(\vec{c})$ is witnessed true in N.

It is the witnessing of $\forall$-clauses that, in essence, enables the implicit learning of quantified generalizations. Let us see how that works. Intuitively, from examples $\phi(\vec{c}_1), \ldots,$ one would like to generalize to $\forall\vec{x}\phi(\vec{x})$, the latter being a statement about infinitely many objects. But what criteria would justify this generalization, outside of (say) witnessing infinitely many instances? Our result shows that, surprisingly, it suffices to get finitely many examples, so as to witness $\phi(\vec{c}_1), \ldots, \phi(\vec{c}_k)$ and yield universally quantified sentences with high probability. This is possible because, via Theorem 2, all the names not mentioned in the KB and the query behave "identically." Thus, provided we witness the grounding of ϕ for a sufficient but finite set of constants, we can treat the implicit KB as including $\forall$-clauses, as it yields the same judgments on our queries.

Putting it all together, formally, in any given learning epoch, let S be the class of queries we are interested in asking: that is, S is any finite set of ground formulas. Let C then be all the names mentioned in S, the KB, and z extra new ones chosen arbitrarily, where z is at least the rank of the KB. If z = KB's rank, then the rank of the implicit KB matches that of the explicit KB; otherwise, it would be higher. So the definition says that the witnessing of $\forall\vec{x}\phi(\vec{x})$ happens when $\phi(\vec{c})$ is witnessed for all $\vec{c} \in C$. We think this notion is particularly powerful, as it neither makes references to bindings from the full set of names $\mathbb{N}$ (which is infinite), nor to not observing negative instances. Note also that witnessing does not require observing all atoms: a clause is witnessed to evaluate to true if some literal appearing in it is true in the partial model. Thus, the $\forall$-clause witnessed may involve predicates not explicitly appearing in the partial model.

Witnessed formulas correspond to the *implicit* KB. In order to capture the inferences that the implicit KB permits, we will use partial models to simplify complex formulas in the KB or query. To that end, we define:

Definition 12: Given a partial model N and a propositional formula ϕ, the *restriction of ϕ under N,* denoted $\phi|_N$, is recursively defined: if ϕ is an atom witnessed in N, then $\phi|_N$ is the value that ϕ is witnessed to evaluate to under N; if ϕ is an atom not set by N, then $\phi|_N = \phi$; if $\phi = \neg\psi$, then $\phi|_N = \neg(\psi|_N)$; and if $\phi = \alpha \wedge \beta$, then $\phi|_N = (\alpha|_N) \wedge (\beta|_N)$. (And analogously for Boolean connectives $\vee$ and $\supset$.) For a partial model N and set of propositional formulas F, we let $F|_N$ denote the set $\{\phi|_N : \phi \in F\}$.

Notice that here we do not define restrictions for quantified formulas, such as those appearning in the KB: while that is possible it is not needed, as we will be leveraging Theorem 2 for reasoning.

5 Implicit Learnability

The central motivation here is learning to reason in FOL, and as argued earlier, implicit learning circumvents the need for an explicit hypothesis, especially since hypothesis fitting is intractable, unless one severly restricts the hypothesis space. So, learning is integrated tightly into the application using the knowledge extracted from data. Our definitions in the previous sections establish the grounds for which a first-order implict KB can be learned from finitely many finite-size examples, but also the grounds for deciding propositional entailments of ∀-clauses specified explicitly – i.e., the background knowledge. (Of course, reasoning is not yet tractable, but simply decidable; we return to this point later). Overall, the learning regime is presented in Algorithm 1, and its correctness is justified in Theorem 13.

Algorithm 1 Reasoning with implicit learning

Input: Partial models $N^{(1)}, N^{(2)}, \ldots, N^{(m)}$, explicit KB Δ, query α (a ground formula), number of names k at least equal to Δ's rank
Output: $\hat{p} \in [0,1]$ estimating α is $\hat{p}$-valid (See Theorem 13)
Initialize $v \leftarrow 0$
for $i = 1, \ldots, m$ **do**
  for all k-tuples of names $(c_1, \ldots, c_k)$ from $N^{(i)}$ *not* appearing in $\Delta \wedge \neg\alpha$ **do**
    if $\mathrm{GND}(\Delta \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}}$ is unsatisfiable **then**
      Increment v and skip to the next i.
    end if
  end for
end for
Return v/m

Theorem 13: *Let $\delta, \gamma \in (0,1)$ and $k \in \mathbb{N}$ be given. Suppose we have m partial models drawn i.i.d. from a common distribution D masked by a masking process Θ, where $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$. (Here,* ln *denotes the natural logarithm.) With probability at least $1-\delta$, Algorithm 1 returns a value $\hat{p}$ s.t.*

I *if $\Delta \supset \alpha$ is at most p-valid, $\hat{p} \leq p + \gamma$*

II *if there is a KB $\mathcal{I}$ such that*

1. *$\Delta \wedge \mathcal{I} \models \alpha$,*
2. *the rank of $\Delta \wedge \mathcal{I}$ is at most k, and*
3. *with probability at least p over partial models $N \in \Theta(D)$, there exists names $c_1, \ldots, c_k$ not appearing in Δ or α, such that every formula in $\mathcal{I}$ is witnessed true in N for $c_1, \ldots, c_k$ together with the names appearing in Δ and α*

then $\hat{p} \geq p - \gamma$.

Proof: Part I: $\hat{p} \leq p + \gamma$ if $\Delta \supset \alpha$ is at most p-valid. We first note that when $\mathrm{GND}(\Delta \wedge \neg\alpha, C)|_{N^{(i)}} \models \bot$ for any set of names C, since $N^{(i)}$ is consistent with the actual model $M^{(i)}$ that produced it, $\mathrm{GND}(\Delta \wedge \neg\alpha, C)|_{M^{(i)}} \models \bot$ as well. Thus, in this case, $\mathrm{GND}(\Delta \wedge \neg\alpha, C)$ is falsified by $M^{(i)}$. Since $|C|$ is at least the rank of Δ, it is easy to see that $\mathrm{GND}(\Delta \wedge \neg\alpha)$, which is logically equivalent to $\Delta \wedge \neg\alpha$, is falsifiable at $M^{(i)}$. So, it must be that the negation of that theory (i.e., $\Delta \supset \alpha$) is satisfied at $M^{(i)}$.

Now, $\Delta \supset \alpha$ is by definition p-valid with respect to this distribution on $M^{(i)}$ if the probability that $\Delta \supset \alpha$ is satisfied by each $M^{(i)}$ is p. Moreover, it follows immediately from Hoeffding's inequality that for $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, the probability that the fraction of times $\Delta \supset \alpha$ is satisfied by $M^{(i)}$ (out of m) exceeds p by more than γ is at most $\delta/2$. Thus, $\hat{p}$, which is at most the fraction of times $\Delta \supset \alpha$ is actually satisfied by $M^{(i)}$, likewise is at most $p + \gamma$ with probability at least $1 - \delta/2$.

Part II: rate of witnessing an implicit KB lower bounds $\hat{p}$. Note that by the grounding trick (Theorem 2), $\Delta \wedge \mathcal{I} \models \alpha$ implies that for any set of names $c_1, \ldots, c_k$ not appearing in Δ or α, $\mathrm{GND}(\Delta \wedge \mathcal{I} \wedge \alpha, \{c_1, \ldots, c_k\}) \models \bot$. Suppose that $\mathcal{I}$ is witnessed true for $c_1, \ldots, c_k$ together with the names in Δ and α in $N^{(i)}$. We note that in the restricted formula $\mathrm{GND}(\Delta \wedge \mathcal{I} \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}}$, the groundings of formulas in $\mathcal{I}$ all simplify to 1 (true), and so $\mathrm{GND}(\Delta \wedge \mathcal{I} \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}} = \mathrm{GND}(\Delta \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}}$. Thus, $\mathrm{GND}(\Delta \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}} \models \bot$, so v is incremented on this iteration. Thus, indeed, $\hat{p} = v/m$ is at least the fraction of times out of m that $\mathcal{I}$ is witnessed true for some set of k names. It again follows from Hoeffding's inequality that for $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, this is at least $p - \gamma$ with probability $1 - \delta/2$.

By a union bound, the two parts hold simultaneously with probability at least $1 - \delta$, as needed. ∎

In essence, the no-overestimation condition is a *soundness* guarantee and the no-underestimation condition is a limited *completeness* guarantee: in other words, if the query logically follows from the explicit KB and examples then the algorithm returns success with an appropriate $\hat{p}$, and vice versa.

6 Tractable Reasoning

Algorithm 1 reduces reasoning with implicit learning to deciding entailment. In order to obtain a tractable algorithm, we generally need to restrict the reasoning task somehow. One approach, taken in the previous work on propositional implicit learning [Juba, 2013], is to "promise" that the query is provable in some low-complexity fragment; for example, it is provable by a small treelike resolution proof (where "small" refers to the number of lines of the proof). Equivalently, we give up on completeness, and only seek completeness with respect to conclusions provable in low complexity in a given fragment. In general, then, one obtains a running time guarantee that is parameterized by the size of the proof of the query. We can take a similar approach here, by using an algorithm for deciding entailment that is efficient when parameterized in such terms. In general, what is needed is a fragment for which we can decide the existence of proofs efficiently, and that is "restriction-closed," meaning that for any partial model N, if we consider the restriction of each line of the proof, we obtain a proof in the same fragment. Most fragments we might consider, including specifically treelike or bounded-width resolution, are restriction-closed. (See [Juba, 2012] for details.)

We will motivate an entirely new strategy here, which offers a semantic perspective to the proof-theoretic view in [Juba, 2013]. One classically sound model-theoretic approach to constraining propositional reasoning is to limit the power of the reasoner, as represented, for example, by the work on tautological entailment [Levesque, 1984]. More recently, [Liu *et al.*, 2004] suggest a simple evaluation scheme for proper$^+$ KBs that gradually increases the power of the reasoner: level 0 is standard database lookup together with unit propagation, level 1 allows for one case split in a clause, level 2 allows two case splits, and so on. The formal intuition is as follows: suppose s is a set of ground clauses and ϕ is a ground query, and let us say its a clause for simplicity. Let $\mathcal{U}(s)$ denote the the closure of s under unit propagation, defined as the least set s' satisfying: (a) $s \subseteq s'$ and (b) if literal $l \in s'$ and $(\neg l \vee c) \in s'$ then $c \in s'$. Then let $\mathcal{V}(s)$ define all possible weakenings: $\{c \mid c \text{ is a ground clause and there is a } c' \in \mathcal{U}(s) \text{ s.t. } c' \subseteq c\}$. Then we define $s \models_z \phi$ (read: "entails at levels z") iff one of the following holds:

- *subsume:* $z = 0$, and $\phi \in \mathcal{V}(s)$;
- *split:* $z > 0$ and there is some clause $c \in s$ such that for all literals $l \in c$, $s \cup \{l\} \models_{(z-1)} \phi$.

This scheme is sound as well as tractable:

Theorem 14: *[Liu* et al.*, 2004] Suppose Δ, ϕ are propositional formulas and $z \in \mathbb{N}$. Then, determining if $\Delta \models_z \phi$ can be done in time $O((|\phi| \times |\Delta|)^{z+1})$. Moreover, if $\Delta \models_z \phi$ then $\Delta \models \phi$.*

We will now see how to leverage these results. First, however, we need the equivalent to restriction-closed, as discussed above.

Proposition 15: *Suppose ϕ, Δ, z are as above. Then if $\Delta \models_z \phi$, and N is any partial model then $(\Delta|_N) \models_z (\phi|_N)$.*

Basically, if ϕ is entailed at level z from Δ, then any restriction of ϕ under N must also be entailed by Δ restricted to N, at least at level z if not lower. Notice that restricting a ground formula is equivalent to simply conjoining the literals true at N with both ϕ and Δ, from which the proof follows. Now, recall from Theorem 2, given a proper$^+$ KB Δ and ground query ϕ, we have $\Delta \models \phi$ iff $\text{GND}^-(\Delta \wedge \neg\alpha)$ is unsatisfiable. Here, since α is already ground, we really only need to make sure that Δ is ground wrt all the names in $\Delta \wedge \neg\alpha$ and k new ones, k being the rank of Δ. So let $\text{GND}^\alpha(\Delta)$ denote precisely such a grounding of Δ. It then follows that $\text{GND}^\alpha(\Delta) \models \alpha$ iff $\text{GND}^-(\Delta \wedge \neg\alpha)$ is unsatisfiable iff $\Delta \models \alpha$. So let Algorithm 1$'$ be exactly like Algorithm 1 except that it accepts a parameter z (for limited reasoning) and replaces the following check:

$\text{GND}(\Delta \wedge \neg\alpha, \{c_1, \ldots, c_k\})|_{N^{(i)}}$ is unsatisfiable **with**

$\text{GND}(\Delta, \{c_1, \ldots, c_k, d_1, \ldots, d_m\})|_{N^{(i)}} \models_z (\alpha|_{N^{(i)}})$, where $\{d_1, \ldots, d_m\}$ is the set of names appearing in α but not in Δ.

Theorem 16: *Let δ, γ, k, m be as in Theorem 13, and let $z \in \mathbb{N}$. Then with a probability at least $1 - \delta$, Algorithm 1$'$ returns a value $\hat{p}$ such that:* **(I)** *and* **(II)** *is as in Theorem 13 except for* **(II.1)** *which states that $\Delta \wedge \mathcal{I} \models_z \alpha$. The algorithm runs in time $O(1/\gamma^2 \times (|\phi| \times |\Delta|)^{z+1} \times log(1/\delta))$.*

Discussion. Interestingly, in [Liu and Levesque, 2005], it is shown that reasoning is also tractable in the first-order case if the knowledge base and the query both use a bounded number of variables. This would then mean that we would no longer be limited to ground queries and can handle queries with quantifiers. This direction is left for future research. Nonetheless, we note that deciding quantified (as opposed to ground) queries appears to demand more from learning. In general, in an infinite domain, we cannot hope to observe in a finite partial model that universally quantified formulas are ever true. Thus, we anticipate that extensions that handle queries with quantifiers will need a substantially different framework, presumably with stronger assumptions. One possible framework takes a more *credulous* approach to the learning problem (in contrast to our *skeptical* approach based on witnessing truth): we suppose that when a formula is frequently false on the distribution of examples, we also frequently obtain a partial model that witnesses the formula false—e.g., a partial model in which a binding of a candidate ∀-clause falsifies it. This is undoubtedly an assumption about the benevolent nature of the environment, captured as the notion of *concealment* in [Michael, 2010], but it does make learning conceptually simpler. In this framework, one permits all conclusions that are not explicitly falsified. Whether such an idea can be used for inductive generalization of FOL formulas over arbitrary distributions remains to be seen.

7 Conclusions

In this work, we presented new results on the problem of answering queries about formulas of first-order logic (FOL) based on background knowledge partially represented explicitly as other formulas, and partially represented as examples independently drawn from a fixed probability distribution. By appealing to the paradigm of implicit learnability, we sidestepped many major negative results, leading to a learning regime that works with a general and expressive FOL fragment. No restrictions were posed on clause length, predicate arity, and other similar technical devices seen in PAC results. Overall, we hope the simplicity of the framework is appealing to the readers and hope our results will renew interest in learnability for expressive languages with quantificational power.

Acknowledgements

B. Juba was supported by NSF Award CCF-1718380. This work was partially performed while B. Juba was visiting the

Simons Institute for the Theory of Computing.

References

[Belle, 2017] V. Belle. Open-universe weighted model counting. In *AAAI*, 2017.

[Billingsley, 1995] P. Billingsley. *Probability and Measure*. Wiley-Interscience, 3 edition, April 1995.

[Cohen and Hirsh, 1994] William W Cohen and Haym Hirsh. The learnability of description logics with equality constraints. *Machine Learning*, 17(2-3):169–199, 1994.

[Cohen and Page, 1995] William W Cohen and C David Page. Polynomial learnability and inductive logic programming: Methods and results. *New Generation Computing*, 13(3-4):369–409, 1995.

[Daniely and Shalev-Shwartz, 2016] Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In *COLT*, pages 815–830, 2016.

[De Raedt and Džeroski, 1994] L. De Raedt and S. Džeroski. First-order *jk*-clausal theories are PAC-learnable. *Artificial Intelligence*, 70(1):375–392, 1994.

[Giacomo *et al.*, 2011] G. De Giacomo, Y. Lespérance, and H. J. Levesque. Efficient reasoning in proper knowledge bases with unknown individuals. In *IJCAI*, pages 827–832, 2011.

[Juba, 2012] Brendan Juba. Learning implicitly in reasoning in pac-semantics. *arXiv preprint arXiv:1209.0056*, 2012.

[Juba, 2013] Brendan Juba. Implicit learning of common sense for reasoning. In *IJCAI*, pages 939–946, 2013.

[Kearns *et al.*, 1994] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[Kersting *et al.*, 2011] K. Kersting, S. Natarajan, and D. Poole. Statistical relational AI: Logic, probability and computation. 2011.

[Khardon and Roth, 1999] Roni Khardon and Dan Roth. Learning to reason with a restricted view. *Machine Learning*, 35(2):95–116, 1999.

[Lakemeyer and Levesque, 2002] G. Lakemeyer and H. J. Levesque. Evaluation-based reasoning with disjunctive information in first-order knowledge bases. In *Proc. KR*, pages 73–81, 2002.

[Levesque and Lakemeyer, 2001] H. J. Levesque and G. Lakemeyer. *The logic of knowledge bases*. The MIT Press, 2001.

[Levesque, 1984] H.J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202. American Association for Artificial Intelligence, 1984.

[Levesque, 1998] H. J. Levesque. A completeness result for reasoning with incomplete first-order knowledge bases. In *Proc. KR*, pages 14–23, 1998.

[Liu and Levesque, 2005] Yongmei Liu and Hector J. Levesque. Tractable reasoning in first-order knowledge bases with disjunctive information. In *Proc. AAAI*, pages 639–644, 2005.

[Liu *et al.*, 2004] Yongmei Liu, Gerhard Lakemeyer, and Hector J. Levesque. A logic of limited belief for reasoning with disjunctive information. In *KR*, pages 587–597, 2004.

[McCarthy and Hayes, 1969] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence*, pages 463–502, 1969.

[Michael, 2009] Loizos Michael. Reading between the lines. In *IJCAI*, 2009.

[Michael, 2010] Loizos Michael. Partial observability and learnability. *Artificial Intelligence*, 174(11):639–669, 2010.

[Moore, 1982] Robert C. Moore. The role of logic in knowledge representation and commonsense reasoning. In *AAAI*, pages 428–433, 1982.

[Muggleton and De Raedt, 1994] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

[Richardson and Domingos, 2006] M. Richardson and P. Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006.

[Srivastava *et al.*, 2014] S. Srivastava, S. J. Russell, P. Ruan, and X. Cheng. First-order open-universe pomdps. In *UAI*, pages 742–751, 2014.

[Valiant, 2000] Leslie G Valiant. Robust logics. *Artificial Intelligence*, 117(2):231–253, 2000.