PROBABILISTIC LOGIC AND INDUCTION

Sebastiaan A. Terwijn *

May 27, 2005

Abstract

We give a probabilistic interpretation of first-order formulas based on Valiants model of pac-learning. We study the resulting notion of probabilistic or approximate truth and take some first steps in developing its model theory. In particular we show that every fixed error parameter determining the precision of universal quantification gives rise to a different class of tautologies. Finally we study the inductive inference of first-order formulas from atomic truths.

1 introduction

The goal of this paper is to develop a notion of model theoretic pac-learning and to study the corresponding notion of probabilistic truth. This parallels the fact that Golds model of language learning [5] can be transformed to a more general model-theoretic one (Osherson et al. [12], see also Terwijn [13]). This has already yielded some interesting results, e.g. connections with the theory of belief revision (Martin and Osherson [11]). The model of pac-learning was introduced by Valiant [15]. This model was the first probabilistic model of learning amenable to a complexity theoretic analysis of learning tasks, and in the subsequent years became one of the most prominent models in the learning theory research. A good introduction to the theory of this model is Kearns and Vazirani [8].

The connections between logic and probability are old and manifold. An early critic of the use of universal statements outside of the synthetic realm of mathematics was the sceptic Sextus Empiricus (2nd–3rd century). He pointed out that without a formal context, where a universal statement can hold by definition, such a statement can only be true when every instance

*Institute of Discrete Mathematics and Geometry, Technical University of Vienna, Wiedner Hauptstrasse 8–10/E104, A-1040 Vienna, Austria, terwijn@logic.at. Supported by the Austrian Research Fund (FWF grant P17503-N12).

of it has been examined. As in most cases this is not feasible, there is in general no way to induce universal statements from examples. As a way out of dilemma's like this, in the course of history there have been many proposals for probabilistic interpretations of quantifiers. In the 20th century, the connections between scientific deduction and the foundations of probability were a major research topic for the members of the Vienna Circle, cf. e.g. Carnap [2]. For a survey of results connecting logic and probability see Halpern [6]. In this paper we give yet another interpretation of universal quantification, based on the theory of pac-learning. Valiant [16] himself already gave an interpretation of logical formula's based on pac-learning. Although some of the ideas in [16] are closely related to the material below, our approach is different. For example, [16] is primarily concerned with a merging of logic and learning, and in particular develops a setting for learning logical rules from statistical data. These rules form a powerful fragment of first order predicate logic, and the learning of a rule consists of producing a good approximation to it. In [16] only finite models are considered. Below we will be using arbitrary first order formulas and we will consider models of arbitrary cardinality.

Another study of probability quantifiers that is related to the logic that we study below is Keisler [9]. The logic $L_{\mathbb{A}P}$ studied there contains a quantifier $(Px \geq r)\varphi$ meaning that the set $\{x : \varphi(x)\}$ has measure at least r. This coincides with our interpretation of universal quantification below. However, negation in $L_{\mathbb{A}P}$ behaves in a crucially different way, and as a consequence $L_{\mathbb{A}P}$ does not contain any of the classical quantifiers $\forall$ and $\exists$, whereas the logic below still contains the classical $\exists$. The same volume in which [9] appeared describes work by H. Friedman on probabilistic quantifiers.

We will not review all studies of probability logic here, but we only mention one other approach, namely the one where instead of models that each have their own probability distribution one considers classical models, with the usual semantics, and where the probability distributions are taken over the class of models. This is the approach studied in Adams [1]. Appendix 7 in [1] contains a brief outline of a theory for predicate logic under this approach. This approach seems to be fundamentally different from the one taken by e.g. Keisler, Valiant, and by us.

Below, we will first give a naive statistical semantics for first order formulas based on sampling according to an unknown distribution $\mathcal{D}$ and an error parameter ε. In Section 3 we discuss the notion of probabilistic or approximate truth resulting from this, and in particular compare it to classical and intuitionistic truth. In Section 4 we then discuss the induction of formulas, which will be the deciding of such formulas with a prescribed rate of certainty. We will not be concerned with beliefs and their relation to probability theory and induction. We refer to Hill, Paris, and Wilmers [7] for interesting results and references on this related topic. A recent survey on logic and learning in artificial intelligence is De Raedt and Kersting [4].

2 A probabilistic interpretation of first-order logic

Fix a language $\mathcal{L}$ of a finite signature. In our setting, there will always be a given model $\mathcal{M}$ and a given probability distribution $\mathcal{D}$ over the universe of $\mathcal{M}$. The idea of the learning situation will be that we have to learn about the structure of the unknown model $\mathcal{M}$ by taking samples from it according to the (also unknown) distribution $\mathcal{D}$. When sampling elements from the model, we will be given the atomic truths these elements satisfy. From this information we have to induce general statements involving full quantification. It is clear that this can only be done with a certain rate of confidence. First we need a definition of approximate truth. The definition of what it means to induce a sentence in this context ("learning") will be given in Section 4. Note that the *language* of our probabilistic logic is just first-order predicate logic, but the interpretation of first-order formulas will be different from the classical one.

Definition 2.1 (Truth definition) Given a first-order sentence φ and $\varepsilon \in [0,1]$, we inductively define $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ as follows.

1. For every prime formula φ (i.e. φ atomic or the negation of an atomic formula), $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ if $\mathcal{M} \models \varphi$.
2. The logical connectives $\wedge$ and $\vee$ are treated classically, e.g. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi \wedge \psi$ if it holds that $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ and $\mathcal{M} \models_{\mathcal{D},\varepsilon} \psi$.
3. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \exists x \varphi(x)$ if $\exists x \in \mathcal{M}\ \ \mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi(x)$.
4. The case of negation is split into subcases as follows:
   - 4.1. For φ atomic, $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\neg\varphi$ if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$. Furthermore, $\neg$ distributes in the classical way over $\vee$ and $\wedge$, e.g. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg(\varphi \wedge \psi)$ if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi \vee \neg\psi$.
   - 4.2. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\exists x \varphi(x)$ if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \forall x \neg\varphi(x)$.
   - 4.3. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\forall x \varphi(x)$ if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \exists x \neg\varphi(x)$.
5. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi \rightarrow \psi$ if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi \vee \psi$.

6. $\mathcal{M} \models_{\mathcal{D},\varepsilon} \forall x \varphi(x)$ if $\Pr_{\mathcal{D}}\big[x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi(x)\big] \geq 1 - \varepsilon$.

Note that in the above definition everything is treated classically, except the interpretation of $\forall x \varphi(x)$ in case 6. Case 4 in the definition allows us to rewrite all formula's in prenex normal form by pushing the negations inside.

Note that both $\mathcal{M} \models_{\mathcal{D},\varepsilon} \exists x \varphi(x)$ and $\mathcal{M} \models_{\mathcal{D},\varepsilon} \forall x \neg\varphi(x)$ may hold, since the interpretation of the first is the classical one, but the interpretation of the second is that *most* x's satisfy $\neg\varphi(x)$. That is, the logic of $\models_{\mathcal{D},\varepsilon}$ is *paraconsistent*.

The asymmetry in the interpretation of $\exists$ and $\forall$ can be explained when one thinks of establishing with a given degree of confidence the truth of statements by taking samples: If the sample contains an x with $\varphi(x)$ one knows with certainty that $\exists x \varphi(x)$, but if one is looking for evidence for the statement $\forall x \varphi(x)$ in general only a certain degree of confidence can be achieved. This interpretation of universal statements is in line with e.g. Popper's philosophy of science, where one counterexample says more than any number of positive examples. (Note though that Popper's philosophy was lacking a proper probabilistic interpretation.)

Note that for $\varepsilon = 0$ the truth definition does not coincide with the classical one, since in this case there can still be exceptions to a universal statement, although they can only form a set of measure zero. Still, the case $\varepsilon = 0$ is somewhat special, as exemplified by Theorem 3.7.2 below.

One could also propose to interpret $\mathcal{M} \models_{\mathcal{D},\varepsilon} \exists x \varphi(x)$ by $\Pr_{\mathcal{D}}\big[x : \mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi(x)\big] > \varepsilon$, so that the cases of $\exists x \neg\varphi(x)$ and $\forall x \varphi(x)$ are exactly complementary, instead of having a small overlap as in the definition above. However, besides the fact that the intended meaning of $\exists$ changes, it then becomes impossible to distinguish between these two cases on the basis of finite samples with a prescribed degree of confidence. Since our definition of learnability (Definition 4.1) will require exactly the making of such a decision, the definition above is more suited for our purposes.

Below, terms like "probabilistic validity" refer to the probabilistic interpretation above. So we will say that a sentence φ is *probabilistically valid* (or a probabilistic tautology) if for all $\mathcal{M}$, $\mathcal{D}$, and ε it holds that $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$. Similarly, φ is *probabilistically satisfiable* if there exist $\mathcal{M}$, $\mathcal{D}$, and ε such that $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$.

Example 2.2 Consider the sentence

$$\varphi \equiv \forall x R(x) \wedge \forall x Q(x) \wedge \forall x \big(R(x) \leftrightarrow \neg Q(x)\big).$$

Then φ is probabilistically satisfiable when $\varepsilon \geq \frac{1}{3}$, but not when $\varepsilon < \frac{1}{3}$. Namely to every one of the three universal statements the exceptions can have at most ε in weight. An optimum is obtained when all exceptions are equal in weight, and then $\varepsilon = \frac{1}{3}$.

3 Probabilistic truth

The next proposition allows us to compare approximate truth with other kinds of logical truth.

Proposition 3.1 *Let $\varphi(x)$ be a formula with a free variable x. Then the sets $Y = \{x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi(x)\}$ and $N = \{x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi(x)\}$ satisfy $Y \cup N = \mathcal{M}$, but not necessarily $Y \cap N = \emptyset$.*

Proof. We prove this by induction on the complexity of φ. The atomic case, the cases of the propositional connectives, and the case of the existential quantifier are all classical, hence trivial. So we only have to consider the case of the universal quantifier. Suppose $\varphi(x) = \forall y \psi(x, y)$, and suppose that x is not in Y. Then $\Pr_{\mathcal{D}}\big[y : \mathcal{M} \models_{\mathcal{D},\varepsilon} \psi(x, y)\big] < 1-\varepsilon$, and by induction hypothesis we have $\Pr_{\mathcal{D}}\big[y : \mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\psi(x, y)\big] \geq \varepsilon$ So certainly there exists $y \in \mathcal{M}$ such that $\neg\varphi(x, y)$. But this is by definition the same as $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\forall y\varphi(x, y)$, so $x \in N$. □

It is instructive to compare the above interpretation of logical formulas to the intuitionistic and the classical one. In classical logic, the interpretations of φ and $\neg\varphi$ are exactly complementary: The sets Y and N from Proposition 3.1 are disjoint and satisfy $Y \cup N = \mathcal{M}$. In intuitionistic logic, Y and N are also disjoint, but they need not satisfy $Y \cup N = \mathcal{M}$. In our present probabilistic interpretation, we do have that $Y \cup N = \mathcal{M}$, but Y and N need *not* be disjoint: there may be an overlap between the interpretations of φ and $\neg\varphi$. In fact, it may happen that $Y \cap N = \mathcal{M}$.

Proposition 3.2 *Every classically satisfiable formula is probabilistically satisfiable, but not vice-versa.*

Proof. Every classically valid/satisfiable formula is also valid/satisfiable under the present probabilistic interpretation: This follows since every formula can be written in prenex normal form (see the remarks following Definition 2.1) and since case 6 in Definition 2.1 is *weaker* than the classical interpretation. In particular we have for any model $\mathcal{M}$ that $\mathcal{M} \models \varphi \Longrightarrow (\forall\mathcal{D}, \varepsilon)\big[\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi\big]$. An example of a probabilistically satisfiable formula that is not classically satisfiable is the sentence $\exists x R(x) \wedge \neg\exists x R(x)$. Note that this sentence is even probabilistically satisfiable with $\varepsilon = 0$. □

Example 3.3 Consider a language $\mathcal{L}$ with predicates $\leq$ and R. Let $\mathsf{lin} = (\forall x, y)[x \leq y \vee y \leq x]$ be the sentence saying that $\leq$ is a linear order and let

$$\varphi = \neg\mathsf{lin} \vee \exists x \forall y (y \leq x).$$

Clearly φ is not a classical tautology. We show that φ probabilistically holds for all $\varepsilon > 0$ in all linear orders that are countable unions of intervals. Let $\mathcal{M}$, $\mathcal{D}$, ε be a probabilistic model with $\varepsilon > 0$. When $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\mathsf{lin}$ then we are done. When $\mathcal{M} \not\models_{\mathcal{D},\varepsilon} \neg\mathsf{lin}$ then we have classically $\mathcal{M} \models \mathsf{lin}$ so $\leq$ really is a linear order in $\mathcal{M}$. Now suppose that $\mathcal{M}$ is a countable union of intervals. Then we can choose $x \in \mathcal{M}$ such that most of the weight is to the left of x: $\Pr_{\mathcal{D}}\big[y \in \mathcal{M} : x \leq y\big] < \varepsilon$. Then in particular

$$\begin{aligned} \Pr_{\mathcal{D}}\big[y : \mathcal{M} \models_{\mathcal{D},\varepsilon} y \leq x\big] &= \\ \Pr_{\mathcal{D}}\big[y : \mathcal{M} \models y \leq x\big] &\geq 1 - \varepsilon \end{aligned}$$

and hence $\mathcal{M} \models_{\mathcal{D},\varepsilon} (\exists x)(\forall y)[y \leq x]$.

Note that this also gives an example of a model $\mathcal{M}$ for which $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ for all $\varepsilon > 0$ does not imply $\mathcal{M} \models_{\mathcal{D},0} \varphi$.

Next we show that φ is not a probabilistic tautology, even when $\varepsilon > 0$. Let $\mathcal{M} = (\omega_1, \leq)$, where $\leq$ is the usual well-order on ω_1. By a 'tail' of ω_1 we understand any set of the form $\big\{y \in \omega_1 : x \leq y\big\}$ for $x \in \omega_1$. Define a measure $\mathcal{D}$ on $\mathcal{M}$ by defining for $A \subseteq \mathcal{M}$,

$$\mathcal{D}(A) = \begin{cases} 1 & \text{if } A \text{ contains a tail of } \omega_1, \\ 0 & \text{if } A \text{ is countable.} \end{cases}$$

It is easy to check that $\mathcal{D}$ is a probability measure. Notice that every initial segment of ω_1 has measure 0. That $\mathcal{D}$ satisfies countable additivity is guaranteed by the fact that the cofinality $\mathrm{cof}(\omega_1)$ of ω_1 is $\aleph_1$ (see Kunen [10]). In particular, however we choose $x \in \omega_1$, all the weight is always to the right of x. Hence $\mathcal{M} \not\models_{\mathcal{D},\varepsilon} \varphi$ for any $\varepsilon > 0$. Uncountable models such as $\mathcal{M}$ in which all weight is "at infinity" will be useful as countermodels in Theorem 3.7.

Lemma 3.4 *Let $\mathcal{D}$ be a probability distribution on $\mathcal{M}$ such that for all $x \in \mathcal{M}$, $\mathcal{D}(\{x\}) \neq 0$. Then for every formula φ, $\mathcal{M} \models \varphi \iff \mathcal{M} \models_{\mathcal{D},0} \varphi$.*

Proof. One direction follows from Proposition 3.2. For the converse direction, if $\mathcal{D}$ is as in the lemma and $\Pr_{\mathcal{D}}\big[x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},0} \varphi(x)\big] = 1$ then in fact $(\forall x \in \mathcal{M})\big[\mathcal{M} \models_{\mathcal{D},0} \varphi\big]$. So the interpretation of $\forall$ is in fact the classical one, and hence every formula is interpreted classically. $\square$

Proposition 3.5 *The probabilistic tautologies coincide with the classical tautologies.*

Proof. That every classically valid formula is also probabilistically valid was proven in Proposition 3.2. For the converse, suppose that φ is not classically valid. Then there is a countable model $\mathcal{M}$ such that $\mathcal{M} \not\models \varphi$. Since $\mathcal{M}$ is countable, there is a distribution $\mathcal{D}$ on $\mathcal{M}$ such that for all $x \in \mathcal{M}$, $\mathcal{D}(\{x\}) \neq 0$. But then by Lemma 3.4, $\mathcal{M} \not\models_{\mathcal{D},0} \varphi$. Hence φ is not probabilistically valid. $\square$

Proposition 3.5 shows that the probabilistic tautologies coincide with the classical tautologies because of the $\varepsilon = 0$ case. Next we refine Proposition 3.5 by considering ε-tautologies for every separate ε.

Definition 3.6 For $\varepsilon \in [0,1]$, a sentence φ an ε-*tautology* if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ for every $\mathcal{M}$ and $\mathcal{D}$.

Theorem 3.7

1. *For all $\varepsilon < \varepsilon'$, the ε-tautologies are included in the ε'-tautologies.*
2. *Although for $\varepsilon = 0$, ε-truth is not the same as classical truth, the* 0*-tautologies coincide with the classical tautologies.*
3. *The class of* 1*-tautologies is different from the ε-tautologies for every $\varepsilon < 1$.*
4. *For all ε, ε' with $1 > \varepsilon' > \varepsilon \geq 0$, the ε-tautologies are different from the ε'-tautologies.*

Proof. Ad 1. This is immediate from Definition 2.1, since case 6 becomes weaker if ε becomes bigger.

Ad 2. This was proved in Proposition 3.5.

Ad 3. This is a degenerate case: Note that for $\varepsilon = 1$ case 6 in Definition 2.1 is in fact empty. This makes a sentence like $\varphi = \forall x R(x)$ a 1-tautology. Clearly φ is not an ε-tautology for any $\varepsilon < 1$.

Ad 4. Since the rationals are dense in the reals it is sufficient to prove this for rational ε and ε'. We use ideas similar to Example 3.3. First we give an example of a $\frac{2}{3}$-tautology φ that is not a ε-tautology for any $\varepsilon < \frac{2}{3}$. Given a unary predicate X, let

$$\mathsf{lin}(X) = \forall x \forall y \big[X(x) \wedge X(y) \rightarrow (xRy \vee yRx)\big]$$

be the sentence saying that R is a linear order on X. For unary predicates X and Y let XRY be the sentence $\forall x \forall y (X(x) \wedge Y(y) \rightarrow xRy)$. Now let X_0, X_1, and X_2 be unary predicates and define

$$\begin{aligned} \mathsf{3lin} = {} & \forall x \big(X_0(x) \vee X_1(x) \vee X_2(x) \big) \wedge \\ & \mathsf{lin}(X_0) \wedge \mathsf{lin}(X_1) \wedge \mathsf{lin}(X_2) \wedge \\ & X_0 R X_1 \wedge X_1 R X_2 \wedge X_2 R X_0. \end{aligned}$$

We do *not* require R to be transitive except on the X_i, e.g. we do not require $X_0 R X_2$. Note that because $\mathsf{3lin}$ is purely universal we have that if $\mathcal{M} \not\models_{\mathcal{D},\varepsilon} \neg\mathsf{3lin}$ then $\mathcal{M}$ consists precisely of X_0, X_1, and X_2 (with no exceptions), the X_i are really (classically) linearly ordered, and X_i precedes X_{i+1} (with indices taken mod 3). Now define[1]

$$\begin{aligned} \varphi = \neg\mathsf{3lin} & \vee \exists x \forall y (X_0(y) \wedge yRx) \\ & \vee \exists x \forall y (X_1(y) \wedge yRx) \\ & \vee \exists x \forall y (X_2(y) \wedge yRx). \end{aligned}$$

We claim that φ is a $\frac{2}{3}$-tautology. Indeed, if $\mathcal{M} \not\models_{\mathcal{D},\varepsilon} \neg\mathsf{3lin}$ then the measure is divided over X_0, X_1, and X_2, so at least one of them has measure at least $\frac{1}{3}$. If this holds for X_i we can pick the upper bound x from X_{i+1}. Next we show that φ is not an ε-tautology for any $\varepsilon < \frac{2}{3}$. Namely let $X_i = \omega_1$ for $i \in \{0, 1, 2\}$, with R on ω_1 the usual well-order, and such that $\mathcal{M}$ consisting of the X_i satisfies $\mathsf{3lin}$. Define a measure on ω_1 by

$$\mathcal{D}(A) = \begin{cases} \frac{1}{3} & \text{if } A \text{ contains a tail of } \omega_1 \\ 0 & \text{if } A \text{ is countable.} \end{cases}$$

$\mathcal{D}$ defines a probability measure on $\mathcal{M}$ by letting $A \subseteq \mathcal{M}$ have measure 0, $\frac{1}{3}$, $\frac{2}{3}$, or 1 depending on whether A contains 0, 1, 2, or 3 tails. (That $\mathcal{D}$ is countably additive again uses that $\mathrm{cof}(\omega_1) = \aleph_1$ as in Example 3.3.) Now any element $x \in \mathcal{M}$ can be R-upper bound for at most one X_i, hence cover at most $\frac{1}{3}$ in measure. This shows that φ is not an ε-tautology for any $\varepsilon < \frac{2}{3}$.

Next we indicate how to generalize the previous construction to obtain a $\frac{1}{3}$-tautology φ that is not an ε-tautology for any $\varepsilon < \frac{1}{3}$. Define the formula

$$\sigma_{ij} = \exists x \exists y \forall z \big((X_i(z) \wedge Rzx) \vee (X_j(z) \wedge Rzx) \big)$$

[1] The simpler sentence $\varphi = \neg\mathsf{3lin} \vee \exists x \forall y (yRx)$ would suffice for this case, but we use the more complex sentence in order to explain how the generalization works.

and define

$$\varphi = \neg \mathsf{3lin} \vee \sigma_{01} \vee \sigma_{02} \vee \sigma_{12}.$$

Basically φ says that if $\mathcal{M} \not\models_{\mathcal{D},\varepsilon} \neg\mathsf{3lin}$ then there are two upper bounds in $\mathcal{M}$ that together cover at least two of the three X_i. Now it is easy to see that there are always two copies X_i that together carry at least $\frac{2}{3}$ of the measure. Hence φ is a $\frac{1}{3}$-tautology. The argument that it is not an ε-tautology for any $\varepsilon < \frac{1}{3}$ is completely analogous to the argument in the previous case.

Now it should be clear how to proceed in the general case of $\varepsilon = 1 - \frac{m}{n}$. In this case we have n copies $X_0, \dots, X_{n-1}$ and a formula $n\mathsf{lin}$ analogous to $\mathsf{3lin}$. For every of the $\binom{n}{m}$ choices $\{i_1, \dots, i_m\}$ of different values from $\{0, \dots, n-1\}$ we have the formula

$$\sigma_{i_1 \dots i_m} = \exists x_1 \dots \exists x_m \forall z \big((X_{i_1}(z) \wedge Rzx) \vee \dots \vee (X_{i_m}(z) \wedge Rzx) \big),$$

and we define

$$\varphi = \neg \mathsf{3lin} \vee \bigvee_{i_1 \dots i_m} \sigma_{i_1 \dots i_m}.$$

The arguments that φ is an $(1 - \frac{m}{n})$-tautology but not an ε-tautology for any $\varepsilon < 1 - \frac{m}{n}$ are completely analogous to the case of $\varepsilon = 1 - \frac{2}{3}$. $\square$

4 Learning logical sentences

Now that we have developed a notion of approximate truth, we want to continue our discussion of inducing general sentences from atomic data. In the definition of learning below there will be given two parameters: An error parameter ε and a confidence parameter δ. Both are typically small numbers from $(0, 1]$. Now given a sentence φ, we say that an algorithm L pac-learns φ if L, given any ε and δ, and with the use of sampling from the unknown model $\mathcal{M}$ according to the also unknown distribution $\mathcal{D}$, can decide the approximate truth of φ in $\mathcal{M}$ (measured using the relation $\models_{\mathcal{D},\varepsilon}$) with high probability, namely $1 - \delta$. Like in Valiants pac-model, the acronym “pac” stands for “probably approximately correct”, where the “probably” refers to the confidence parameter δ and “approximately correct” to the error parameter ε.

In the next definition we make use of a *sampling oracle* EX($\mathcal{D}$), which when called upon randomly draws an element x from the model $\mathcal{M}$, according to the distribution $\mathcal{D}$. Given a sample of elements, the oracle supplies us with all the atomic truths these elements satisfy, for every relation of every arity in the language $\mathcal{L}$. Recall that we have assumed that the language $\mathcal{L}$ is of finite signature, so that every sample satisfies only finitely many atomic truths.

Definition 4.1 (Probabilistic induction) A (probabilistic) algorithm L pac-learns sentence φ if L, for any unknown $\mathcal{M}$ and $\mathcal{D}$, given error parameter $\varepsilon > 0$ and confidence parameter $\delta > 0$, and with access to the sampling oracle $\mathrm{EX}(\mathcal{D})$, L outputs one of the possibilities $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$, $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi$, such that with probability at least $1-\delta$ the output is correct. Note that *both* possible outputs can be correct (see the discussion following Definition 2.1).

For the next result we will assume that $\mathcal{L}$ has no constants or function symbols. See however the remarks in Section 5 that show that this restriction is rather immaterial.

Theorem 4.2 *There exists an algorithm L that pac-learns any sentence φ. If φ has n quantifiers, L takes a sample of size $\left(\frac{1}{\varepsilon}\ln\frac{n}{\delta}\right)^n$.*

Proof. The idea is that one can decide universal quantifiers with any prescribed accuracy by taking large enough samples, and likewise decide existential quantifiers by searching examples in large enough samples. When one knows the number of quantifiers one can iterate this, and compute the size of the sample needed to get a good answer with high probability. More precisely, let φ be any sentence. As before, we may assume that φ is in prenex normal form: $\varphi = \exists x_0 \forall x_1 \dots Q x_n R(x_0,\dots,x_n)$. Let $m \in \omega$. Consider the following induction procedure L: For every x_i sample m x_{i+1}'s from $\mathcal{M}$ according to $\mathcal{D}$. So in total L takes a sample of size m^n.
Claim: If $m > \frac{1}{\varepsilon}\ln\frac{1}{\delta}$ then with certainty $(1-\delta)^n$ one can decide whether $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ or $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi$.
We prove the claim by induction on the number of quantifiers n:

$n = 0$. The base of the induction is in fact empty since we have assumed that the language $\mathcal{L}$ has no constants and function symbols, so that there are no sentences without quantifiers.

$n+1$. Keeping the notation from above, write $\varphi = \exists x_0 \psi(x_0)$.

- Suppose the sample S taken by L satisfies φ (where the quantifiers are restricted to S). We denote this by φ^S. That means that S contains x_0 such that S satisfies $\psi(x_0)$. By induction hypothesis this then holds with probability $\geq (1-\delta)^{n-1}$. Hence $\mathcal{M} \models_{\mathcal{D},\varepsilon} \varphi$ with probability $\geq (1-\delta)^{n-1} \geq (1-\delta)^n$.

– Suppose the sample S does not satisfy φ: No x_0 is found. Suppose that

$$\Pr_{\mathcal{D}}\big[x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\psi(x)\big] < 1 - \varepsilon.$$

Then by Proposition 3.1 we have that

$$\Pr_{\mathcal{D}}\big[x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \psi(x)\big] \geq \varepsilon.$$

Then the probability that the m sampled x's miss this set is $(1-\varepsilon)^m \leq e^{-m\varepsilon} < \delta$ when $m > \frac{1}{\varepsilon}\ln\frac{1}{\delta}$. By induction hypothesis we have that for every x with $\mathcal{M} \models_{\mathcal{D},\varepsilon} \psi(x)$ that $\psi^S(x)$ holds with probability at least $(1-\delta)^{n-1}$. So the probability that φ^S is at least $(1-\delta)(1-\delta)^{n-1}$. So with probability at least $(1-\delta)^n$ it holds that

$$\Pr_{\mathcal{D}}\big[x \in \mathcal{M} : \mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\psi(x)\big] \geq 1 - \varepsilon,$$

i.e. that $\mathcal{M} \models_{\mathcal{D},\varepsilon} \neg\varphi$.

Now to finish the proof of the theorem we notice that if we replace δ in the above claim by $\frac{\delta}{n}$ then we can decide φ with certainty $\big(1-\frac{\delta}{n}\big)^n$. From the binomial or Taylor expansion of the latter expression one can see that $\big(1-\frac{\delta}{n}\big)^n \geq 1-\delta$ for all $\delta \in [0,1]$ and all n, so the theorem follows. □

5 Concluding remarks

- A possible objection to the probabilistic model ω_1 with the measure $\mathcal{D}$ as defined in Example 3.3 is that the relation $\leq$ is not a $\mathcal{D}^2$-measurable subset of ${\omega_1}^2$. (This is a famous argument of Sierpinksi using Fubini's theorem.) Case 6 in Definition 2.1 does not require that all k-ary relations occurring in φ are $\mathcal{D}^k$-measurable in $\mathcal{M}^k$, only that the appropriate sections are $\mathcal{D}$-measurable, but one could argue that the cases where the whole relations are not $\mathcal{D}^k$-measurable are pathological. A natural extra condition would be to require that

  for every k-ary predicate R occurring in φ the set of k-tuples satisfying R is $\mathcal{D}^k$-measurable, (1)

  where $\mathcal{D}^k$ denotes the product measure on $\mathcal{M}^k$. If we wish to impose condition (1) then we have to reprove Theorem 3.7 using other countermodels than the ones used there. This can indeed be done, so that Theorem 3.7 remains true also with condition (1). Note also that we do *not* have (as was already remarked in Keisler [9]; the same argument works here) that $\mathcal{M} \models_{\mathcal{D},\varepsilon} \forall x \forall y R(x,y)$ if and only if $\mathcal{M} \models_{\mathcal{D},\varepsilon} \forall y \forall x R(x,y)$, even under the measurability condition (1). (There exists an easy 3-element counterexample for $\mathcal{M}$.)

- In the proof of Theorem 4.2 we assumed that the language $\mathcal{L}$ had no constants and function symbols. However, if we assume that the truth of all atomic sentences in $\mathcal{L}$ is given (these are only finitely many since $\mathcal{L}$ is of finite signature) then Theorem 4.2 still holds for languages with constants and function symbols. That the truth of all finitely many atomic sentences from $\mathcal{L}$ is given could be accounted for by broadly interpreting the action of the sampling oracle as listing not only all atomic truths of the elements of a sample but also that of all atomic sentences built from constants.

- The setting of Section 4 can be suitably generalized by allowing the possibility that in inducing a formula φ different distributions $\mathcal{D}$ are used for different variables and predicates in φ, instead of the same one for all.

- We note that Definition 4.1 does not subsume the definition of pac-learning, primarily because the task of deciding the truth of a formula (as in Definition 4.1) is different from producing a concept hypothesis from a concept class (as in pac-learning), even if these concepts are defined by first order formulas.

- We have left many model-theoretic questions about the probabilistic logic studied here unsettled. Some of these questions, as well as matters of decidability, are treated in [14].

References

[1] E. W. Adams, *A primer of probability logic*, CSLI Publications, Stanford, 1998.

[2] R. Carnap, *The logical foundations of probability*, University of Chicago Press, 2nd edition, 1963.

[3] C. C. Chang and H. J. Keisler, *Model theory*, North-Holland, 1973.

[4] L. De Raedt and K. Kersting, *Probabilistic logic learning*, SIGKDD Explorer Newsletter 5(1) (2003) 31–48.

[5] E. M. Gold. *Language identification in the limit*, Information and Control 10 (1967) 447–474.

[6] J. Y. Halpern, *An analysis of first-order logics of probability*, Proceedings of 11th International Joint Conference on Artificial Intelligence (IJCAI-89) (1989) 1375–1381.

[7] M. J. Hill, J. B. Paris, and G. M. Wilmers, *Some observations on induction in predicated probabilistic reasoning*, Journal of Philosophical Logic 31 (2002) 43–75.

[8] M. J. Kearns, U. V. Vazirani, *An introduction to computational learning theory*, MIT Press, 1994.

[9] H. J. Keisler, *Probability quantifiers*, in: J. Barwise and S. Feferman (eds.), Model-Theoretic Logics, Springer-Verlag 1985, 509–556.

[10] K. Kunen, *Set Theory: An Introduction to Independence Proofs*, North Holland, 1983.

[11] E. Martin and D. Osherson, *Scientific discovery based on belief revision*, Journal of Symbolic Logic 62(4) (1997) 1352–1370.

[12] D. N. Osherson, M. Stob, S. Weinstein. *A universal inductive inference machine*, Journal of Symbolic Logic 56 (2) (1991) 661–672.

[13] S. A. Terwijn, *Learning and computing in the limit*, to appear in proceedings of the Logic Colloquium 2002.

[14] S. A. Terwijn, *Decidability and undecidability in probabilistic logic*, in preparation.

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

[16] L. G. Valiant, *Robust logics*, Artificial Intelligence 117 (2000) 231–253.