Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Theoretical Computer Science 299 (2003) 37–63

Theoretical Computer Science

www.elsevier.com/locate/tcs

PAC learning of probability distributions over a discrete domain

Letizia Magnoni *, Massimo Mirolli, Franco Montagna, Giulia Simi

Dipartimento di Matematica, Via del Capitano 15, 53100 Siena, Italy

Received March 1999; received in revised form November 2000; accepted May 2001
Communicated by M. Nivat

Abstract

We investigate learning of classes of distributions over a discrete domain in a PAC context. We introduce two paradigms of PAC learning, namely absolute PAC learning, which is independent of the representation of the class of hypotheses, and PAC learning wrt the indexes, which heavily depends on such representations. We characterize non-computable learnability in both contexts. Then we investigate efficient learning strategies which are simulated by a polynomial-time Turing machine. One strategy is the frequentist one. According to this strategy, the learner conjectures a hypothesis which is as close as possible to the distribution given by the frequency relative to the examples. We characterize the classes of distributions which are absolutely PAC learnable by means of this strategy, and we relate frequentist learning wrt the indexes to the $NP = RP$ problem. Finally, we present another strategy for learning wrt the indexes, namely learning by tests.

1. Introduction

Learning Theory deals with the problem of detecting the general principles governing a phenomenon by means of examples of it. The source of the general theory is Gold's seminal paper [5]. The book [13] contains a general treatment of identification of formal languages (r.e. sets) both on positive and negative examples. Identification of recursive functions is treated e.g. in [12]. Other settings of Learning Theory, like probabilistic learning, are treated e.g. in [2, 14, 15, 9]. The main paradigm is the identification *in the limit* of the target concept on all (or, in probabilistic learning, on sufficiently many) infinite lists of examples, without any reference to either the computational complexity of the learning algorithm or to the number of examples needed to learn.

* Corresponding author.
E-mail address: magnoni@unisi.it (L. Magnoni).

0304-3975/03/$ - see front matter
PII: S0304-3975(01)00265-1

An alternative line of research in Learning Theory, the so called (probabilistic approximately correct (PAC) learning (cf. [16], or [8]) arises from three considerations: (1) in many circumstances, it is not important to learn the concept exactly, a good approximation is sufficient; (2) like in probabilistic learning, one does not require an algorithm that succeeds on *all* lists of examples, but only one that succeeds with high probability; and (3) in general, it is important to learn after a short number of examples. Normally, there are requirements on the learning algorithm; the most common is that such an algorithm has to work in polynomial time.

The present paper deals with learning probability distributions over a countable set in a PAC context. This subject is not completely new: [2] treats probabilistic identification in the limit of classes of distributions. The paper [7] deals with PAC learning of probabilistic functions instead of sets. The paper [3] deals with PAC identification of probability distributions over the reals by means of their density levels (in other words, the learner has to identify the density levels instead of the distribution directly). Finally, [6] deals with PAC learning of distributions over $\{0,1\}^n$, wrt a measure of the error which is motivated by information-theoretic considerations.

In the present paper, we introduce two paradigms of PAC learning. In both paradigms, the learner is supplied with a class $\mathscr{D}$ of distributions over a countable set $S=\{E_n : n\in\mathbf{N}\}$ of events such that at each stage exactly one event of S occurs (for simplicity we can assume that S is the set $\mathbf{N}$ of natural numbers, in other words we think of $n\in\mathbf{N}$ as a code of the event $E_n\in S$) and with a class $\mathscr{R}$ of (representations of) hypotheses; a distribution $d\in\mathscr{D}$ is chosen, and a sequence of examples (coded by natural numbers) is drawn at random according to the distribution d. The learner is also supplied with two positive real numbers, ε and δ called *accuracy* and *confidence*, respectively. The learner has to guess a representation $r\in\mathscr{R}$ for a distribution $d'\in\mathscr{D}$ which is *sufficiently close to* d. In our context, the sentence d *is sufficiently close to* d' means that the distance $D(d,d')=\sum_{n\in\mathbf{N}}|d(n)-d'(n)|$ has to be less than ε. In this paper we assume for simplicity $\varepsilon=\delta=1/n$, where n is a positive natural number. This restriction will not affect the PAC learnability or non-learnability of any class of distributions.

In our opinion, $D(d,d')$ constitutes a natural measure of the error occurring when we guess d' instead of d. The intutitive meaning of $D(d,d')$ is explained by means of the following game: a natural number will be drawn at random according to the distribution d. For every $n\in\mathbf{N}$, let E_n denote the event n *will be drawn*, and F_n denote the event n *will not be drawn*. For *every* $n\in\mathbf{N}$, Palayer II has to bet either on E_n or on F_n. For every n, if Player II chooses E_n, he gets $1-d'(n)$ dollars from Player I if E_n occurs, and gives $d'(n)$ dollars to Player I otherwise; if Player II chooses F_n, he gets $d'(n)$ dollars from Player I if F_n occurs, and pays $1-d'(n)$ dollars to Player I otherwise. Then the distance $D(d,d')$ represents the expectation of the random variable *total income of Player II* when he plays according to his best strategy. The distance D' defined by $D'(d,d')=\sup_{n\in\mathbf{N}}|d(n)-d'(n)|$ seems less natural: for example, for every $n\in\mathbf{N}$, there are distributions d,d' such that $D'(d,d')<1/(n+1)$, and the probabilities of choosing an even number according to the distributions d and d' are respectively, 0 and 1.

The first paradigm taken into consideration, named *absolute PAC learning*, is independent of the representation of the class to be learned. We fix once and for all a countable class $\mathscr{D}_0$ such that for every distribution d there are distributions in $\mathscr{D}_0$ which are arbitrarily close to d. A natural candidate for the role of $\mathscr{D}_0$ is the class $\mathscr{D}_{\text{fin}}$ of *finite distributions*, i.e. the class of rational-valued distributions which are null almost everywhere. We fix once and for all a class R_{fin} of standard representations of distributions of $\mathscr{D}_{\text{fin}}$. The criterion of success is the following: the learner Φ is successful if there is a polynomial $p(x)$ such that for every positive natural number n, for all $d \in \mathscr{D}$, upon seeing $m \geqslant p(n)$ examples drawn at random according to the distribution d, Φ conjectures, with probability $\alpha > 1 - 1/n$, a representation $r \in R_{\text{fin}}$ of a hypothesis d' such that $D(d, d') < 1/n$.

It is worth noticing that in the absolute PAC learning:

(a) The space of hypotheses need not include the class of distributions $\mathscr{D}$ to be learned, it is only required that for every distribution $d \in \mathscr{D}$ there are hypotheses arbitrarily close to d.
(b) The class of hypotheses need not be included in the class $\mathscr{D}$ to be learned, in other words, the learner can conjecture distributions which are not in $\mathscr{D}$.
(c) The number of examples needed by the learner must be independent of the target distribution.

We will prove that if a learner PAC learns a class $\mathscr{D}_0$ of distributions, he also PAC learns its closure (wrt to the metric D). So, it is sufficient to assume that the space of hypotheses contains a dense subset of the class of distributions to be learned.

Our second paradigm, called PAC learning *wrt the indexes*, strongly depends on the class of representations of the distributions of the class $\mathscr{D}$ to be learned. First of all, we require that $\mathscr{D}$ is countable, that the space of hypotheses $\mathscr{R}$ is a set of representations of *exactly* those distributions which are in $\mathscr{D}$, and that the learner has always to guess an hypothesis from $\mathscr{R}$.

The criterion of success is almost as in the case of absolute learning, with the only exception that the number of examples needed to learn depends not only on the accuracy and confidence parameters, but also on the shortest representation of the target distribution. More precisely, the learner Φ is successful if there is a polynomial $p(x, y)$ such that the following holds:

Let d be any distribution of $\mathscr{D}$, let r be the shortest representation of d in $\mathscr{R}$, let $n \in \mathbf{N}$, $n > 0$, and let $m \geqslant p(n, l(r))$, where $l(r)$ denotes the binary length of r. Then, with probability $\alpha > 1 - 1/n$, on a sample of m examples drawn at random according to the distribution d, Φ conjectures a hypothesis $r \in \mathscr{R}$ for a distribution $d' \in \mathscr{D}$ such that $D(d, d') < 1/n$.

We also consider a weaker form of PAC learning wrt the indexes, called *weak* PAC learning, in which the length $l(r)$ of the shortest representation of the target distribution is replaced by the Gödel number of r.

In the learning wrt the indexes, the learner can take advantage of the fact that if the target distribution has only long indexes, then he can use more examples to learn. To the contrary, he has the disadvantage that he must always conjecture hypotheses which

are (representations of) distributions in the class to be learned, whereas in the absolute PAC learning the learner may have a larger class of hypotheses at his disposal.

In Learning Theory, the learner can meet two kinds of obstacles: *structural* obstacles, due to the structure of the class to be learned: oracles cannot help to overcome these obstacles; and *computational* obstacles, due to the fact that learning is potentially possible, but it is impossible to find an efficient learning algorithm, or, in some cases, learning is not algorithmic at all. When learning languages from positive data, we meet a structural obstacle with the class of all finite languages plus the language consisting of all words. This class cannot be learned from positive data, not even by a non-computable learner. To the contrary, the class of languages which are the graphs of single-valued total recursive functions cannot be learned because of computational obstacles: there is a potential learner who learns the class, but such a learner cannot be simulated by a Turing machine.

Now, when learning classes of distributions wrt the indexes, we do not meet structural obstacles: indeed, in Section 3, we show that any countable class of distributions can be PAC learned wrt the indexes by a (possibly non-computable) learner. Moreover, in Section 5, we show that the class of finite distributions is PAC learnable wrt the indexes by means of a P-time learning algorithm. To the contrary, there are computational obstacles (in Section 5): e.g, if $NP \neq RP$, then there is a class of distributions that is absolutely PAC learnable using a P-time algorithm, but cannot be learned wrt to the indexes (according to a suitable representation) by any P-time algorithm, not even using coins.

The situation is different when dealing with absolute PAC learning: in Section 3 we find a characterization of classes of distributions which are absolutely PAC learnable (possibly by a non-computable learner), and we show that the class of uniform distributions over finite sets of natural numbers is not absolutely PAC learnable, not even by a non-computable learner. Thus, there are also structural obstacles to absolute PAC learning.

In Section 4 of the paper we examine some particular learning strategies. The first one works in the case of absolute PAC learning, and is suggested by the law of large numbers. The strategy is the following: on a sample σ of random examples, the learner conjectures the finite distribution f_σ such that for every $n \in \mathbf{N}$, $f_\sigma(n)$ is the frequency of the number n relative to σ. Such a learner is called *the frequentist learner*. We characterize the classes of distributions that are learned by the frequentist learner. These are those classes $\mathscr{D}$ for which there is a polynomial $p(x)$ such that, for all $d \in \mathscr{D}$ and for every positive $n \in \mathbf{N}$ there is a set $A_{d,n} \subseteq \mathbf{N}$ of cardinality at most $p(n)$ such that $\sum_{i \in A_{d,n}} d(i) \geqslant 1 - 1/n$.

Another approach, which concerns learning wrt the indexes, is named *learning by tests* (in Section 6). Roughly, the idea about this strategy is the following: assume that, for each distribution d in the class $\mathscr{D}$ to be learned, we can approximately compute the probabilities (wrt d) of a finite number of sets $C_1 \cdots C_n$. Let finite sequence σ of examples be given. Say that σ *passes the test for* d iff for $i = 1 \cdots n$ the frequency of C_i relative to σ is sufficiently close to the probability C_i wrt d (we will make the

words *sufficiently close* precise in the sequel). The learner conjectures the first index r less than the length of σ for a distribution $d \in \mathscr{D}$ such that σ passes the test for d. This strategy is successful in the case of *weak* PAC learning wrt to the indexes, provided that it is possible to distinguish the distributions of $\mathscr{D}$ by means of these tests, i.e., whenever the same test can fit only with distributions which are very close each other. A similar approach for computable PAC learning wrt the indexes in the stronger sense works if and only if $RP = NP$.

2. Preliminaries

Throughout the whole paper, $\mathbf{N}$, $\mathbf{R}$, $\mathbf{R}^+$, $\mathbf{Q}$ and $\mathbf{Q}^+$ denote, respectively the set of natural numbers, the set of real numbers, the set of non-negative real numbers, the set of rational numbers, and the set of non-negative rational numbers. *Seq* denotes the set of finite sequences of natural numbers, *Bseq* denotes the set of finite binary sequences, and $\mathscr{P}^{<\omega}(\mathbf{N})$ denotes the family of finite subsets of $\mathbf{N}$. The symbol $\emptyset$ denotes the empty set (hence, the empty sequence).

The length of a (finite) sequence τ is denoted by $lth(\tau)$. If $\sigma \in Seq$, and $i < lth(\sigma)$, σ_i denotes the ith element of σ. The sequence whose elements are $a_1, \ldots, a_n$ each in the order given, is denoted by $\langle a_1, \ldots, a_n \rangle$. The symbol $*$ denotes concatenation of sequences. Note that for every $k \in \mathbf{N}$, the set $\{\sigma \in Seq : lth(\sigma) = k\}$ coincides with $\mathbf{N}^k$.

For all $n \in \mathbf{N}$, $l(n)$ denotes the length of the binary expansion of n. We fix once and for all a polynomial time bijection $^\sharp$ from *Seq* onto $\mathbf{N}$ whose inverse is still polynomial time, and such that for $\sigma \in Seq$, $l(\sigma^\sharp)$ is linear in the sum $\sum_{i<lth(\sigma)} l(\sigma_i)$.

A distribution on $\mathbf{N}$ (for short: a *distribution*) is a map d from $\mathbf{N}$ into $\mathbf{R}^+$ such that $\sum_{n\in\mathbf{N}} d(n) = 1$. A distribution is said to be *finite* iff $range(d) \subseteq \mathbf{Q}^+$, and $d(n) = 0$ for almost all n. The set of all distributions is denoted $\mathscr{D}^*$, whereas the set of all finite distributions is denoted $\mathscr{D}_{\text{fin}}$.

If d is a distribution and $\sigma \in Seq$, we set $d(\sigma) = \prod_{i<lth(\sigma)} d(\sigma_i)$. If S is a set of sequences of same length k, we put $d(S) = \sum_{\sigma\in S} d(\sigma)$. It is clear that $d(\sigma)$ denotes the probability of drawing σ with $lth(\sigma)$ draws according to the distribution d, while $d(S)$ denotes the probability of choosing a $\sigma \in S$ with k random draws according to the distribution d. For all non-empty $\sigma \in Seq$, f_σ denotes the finite distribution defined by

$$f_\sigma(n) = \frac{Card\{i < lth(\sigma) : \sigma_i = n\}}{lth(\sigma)}.$$

The distribution f_σ is called *the frequency distribution relative to* σ.

Throughout the whole paper, $\ln(x)$ and $\log(x)$ denote the base e logarithm of x, and the base 2 logarithm of x, respectively.

For typographic reasons we write $\exp(x)$ for e^x. We often use the following.

Proposition 2.1. *The following inequalities hold*:
(a) $1 + x \leqslant \exp(x)$.
(b) *If* $-\frac{1}{2} \leqslant x \leqslant 0$, *then* $1 + x \geqslant \exp(2x)$.

The probability of the event E is denoted by $Pr(E)$. Throughout the paper, we make often use of the following well-known result.

Proposition 2.2 (Chernoff's bounds). *Let $X_1,\ldots,X_m$ denote m independent random variables such that for $i=1\cdots m$, $Pr(X_i=1)=p$, and $Pr(X_i=0)=1-p$. Let $S_m=X_1+\cdots+X_m$. Then for $0\leqslant\gamma\leqslant 1$, the following conditions hold*:

- $Pr(S_m>mp+m\gamma)\leqslant \exp(-\gamma^2 m)$,
- $Pr(S_m<mp-m\gamma)\leqslant \exp(-\gamma^2 m)$,
- $Pr(|S_m-mp|>m\gamma)\leqslant 2\exp(-\gamma^2 m)$.

As a consequence we obtain that if d is any distribution and $E\subseteq \mathbf{N}$, then $Pr(|f_\sigma(E)-d(E)|)>\gamma)\leqslant 2\exp(-\gamma^2 lth(\sigma))$.

Definition 2.3. Let $\mathscr{D}$ be a countable class of distributions. A representation of $\mathscr{D}$ is a pair $\mathscr{R}=\langle R,h\rangle$ where R is a set of finite binary strings, and h is a function from R onto $\mathscr{D}$.

If $\langle R,h\rangle$ is a representation of a countable class $\mathscr{D}$ of distributions, and $r\in R$, we say that r is a representation of $h(r)$.

Definition 2.4. A P-representation of $\mathscr{D}$ is a representation $\mathscr{R}=\langle R,h\rangle$ of $\mathscr{D}$ such that

1. R is in P.
2. There is a P-time function $g(r,n,x):R\times\mathbf{N}\times\mathbf{N}\xrightarrow{g}\mathbf{Q}^+$ such that for all $r\in R$ and for all $n,x\in\mathbf{N}$, $n>0$, one has: $|h(r)(x)-g(r,n,x)|<1/n$.

Definition 2.5. Two representations $\mathscr{R}_1=\langle R_1,h_1\rangle$, $\mathscr{R}_2=\langle R_2,h_2\rangle$ of $\mathscr{D}$ are said to be P-equivalent iff there are P-time computable functions f and g from *Bseq* into *Bseq* which map R_1 into R_2 and R_2 into R_1, respectively, such that $h_2\circ f_{|R_1}=h_1$, and $h_1\circ g_{|R_2}=h_2$.

The next lemma shows that, modulo P-equivalence, we can replace any representation $\mathscr{R}=\langle R,h\rangle$, where R is in P, by one of the form $\mathscr{R}'=\langle\mathbf{N},h'\rangle$, where $\mathbf{N}$ is the set of (binary representations of) natural numbers.

Lemma 2.6. *Let $\mathscr{R}=\langle R,h\rangle$ be a representation of a countable class $\mathscr{D}$ of distributions such that R is in P. Then there is a representation of the form $\mathscr{R}'=\langle\mathbf{N},h'\rangle$ such that $\mathscr{R}$ and $\mathscr{R}'$ are P-equivalent* (*thus in particular, if $\mathscr{R}$ is a P-representation, then $\mathscr{R}'$ also is*).

Proof. Let $^\circ$ be a P-time bijection from *Bseq* onto $\mathbf{N}$ such that its inverse $^{\circ^{-1}}$ is in turn P-time. Fix $\bar{r}\in R$. Define, for $n\in\mathbf{N}$,

$$h'(n)=\begin{cases} h(n^{\circ^{-1}}) & \text{if } n^{\circ^{-1}}\in R,\\ h(\bar{r}) & \text{otherwise.}\end{cases}$$

Now let, for $\tau \in Bseq$, $f(\tau) = \tau^{\circ}$.

Also let, for $n \in \mathbf{N}$,

$$g(n) = \begin{cases} n^{\circ^{-1}} & \text{if } n^{\circ^{-1}} \in R, \\ \bar{r} & \text{otherwise.} \end{cases}$$

Finally, define $g(\tau) = \bar{r}$ for those binary sequences τ that are not the binary expansion of any number.

Clearly, f and g are P-time, and $h' \circ f_{|R} = h$, $h \circ g_{|\mathbf{N}} = h'$. □

Definition 2.7. We define, for $d_1, d_2 \in \mathscr{D}^*$, $D(d_1, d_2) = \sum_{x \in \mathbf{N}} |d_1(x) - d_2(x)|$.

It is readily seen that D is a metric over the space $\mathscr{D}^*$ of all distributions.

Notation 1. For all $\mathscr{D}_0 \subseteq \mathscr{D}^*$, $\bar{\mathscr{D}}_0$ denotes the closure of $\mathscr{D}_0$ wrt the topology induced by the metric D.

Note that $\bar{\mathscr{D}}_{\text{fin}} = \mathscr{D}^*$.

We are going to define some paradigms of PAC learning for classes of distributions. In the tradition of PAC learning, people distinguish between the accuracy parameter ε and the confidence parameter δ. This distinction is relevant if one is looking for optimal bounds to the number of examples needed for learning. However, if one is only interested in learnability-non-learnability in a polynomial number of examples, one can safely *assume that the confidence and the accuracy parameters coincide*. This we will do in the sequel.

In order to define the first paradigm, namely absolute PAC learning, we need a preliminary definition.

Definition 2.8. Fix a P-time bijection $^{-}$ from $\mathbf{Q}^+$ onto $\mathbf{N}$, whose inverse is in turn P-time. The canonical representation $\mathscr{R}_{\text{fin}}$ of the class $\mathscr{D}_{\text{fin}}$ of finite distributions is the unique representation $\langle R_{\text{fin}}, h_{\text{fin}} \rangle$ of $\mathscr{D}_{\text{fin}}$ such that if $d \in \mathscr{D}_{\text{fin}}$, if $d(x) = 0$ for all $x \notin \{a_0 \cdots a_n\}$, where $a_0 < \cdots < a_n$, and for $i = 0 \cdots n$, $d(a_i) = q_i \in \mathbf{Q}^+ - \{0\}$, then the unique representation of d in $\mathscr{R}_{\text{fin}}$ is given by the binary string of $r = (\langle (\langle a_0, \bar{q_0} \rangle)^{\sharp} \cdots (\langle a_n, \bar{q_n} \rangle)^{\sharp} \rangle)^{\sharp}$.

Definition 2.9. Let $\mathscr{D}$ be a class of distributions. We say that $\mathscr{D}$ is absolutely PAC learnable if there are a function $\Phi(\sigma, n) : Seq \times \mathbf{N} \xrightarrow{\Phi} R_{\text{fin}}$ and a polynomial $p(x)$ such that for all $d \in \mathscr{D}$, for all $n > 0$, and for all $k \geqslant p(n)$, one has

$$d\left(\left\{\sigma \in \mathbf{N}^k : D(h_{\text{fin}}(\Phi(\sigma, n)), d) \geqslant \frac{1}{n}\right\}\right) < \frac{1}{n}.$$

In words: if one chooses a sequence σ of $k \geqslant p(n)$ natural numbers at random according to the distribution d, the probability that Φ on σ, n conjectures a d' such that $D(d,d') \geqslant 1/n$ is $< 1/n$.

Definition 2.10. $\mathscr{D}$ is computably absolutely PAC learnable if the conditions of Defintion 2.9 are satisfied, and, in addition, Φ is P-time computable.

This paradigm is called absolute PAC learning because it only depends on the class of distributions and not on the indexing. Note that, since $\bar{\mathscr{D}}_{\text{fin}} = \mathscr{D}^*$, for any distribution d in the class $\mathscr{D}$ to be learned we can find $d' \in \mathscr{D}_{\text{fin}}$ arbitrarily close to d. The next theorem shows that this is exactly what is needed for PAC learning. In other words, in order to PAC learn a class $\mathscr{D}$ of distributions (according to either Definition 2.9 or 2.10) it is sufficient to PAC learn a dense subset of $\mathscr{D}$.

Theorem 2.11. *If Φ be an algorithm of absolute PAC learning for a class $\mathscr{D}$ of distributions, then Φ is an algorithm of PAC learning for $\bar{\mathscr{D}}$.*

Proof. We start from the following Lemma.

Lemma 2.12. *If $D(d,d_0) < \varepsilon$, then $\sum_{\sigma \in \mathbf{N}^k} |d(\sigma) - d_0(\sigma)| < \varepsilon k$.*

Proof. Induction on k. If $k = 1$ the claim is easy: $\sum_{\sigma \in \mathbf{N}^1} |d(\sigma) - d_0(\sigma)| = \sum_{x \in \mathbf{N}} |d(x) - d_0(x)| < \varepsilon$.

Now suppose $\sum_{\sigma \in \mathbf{N}^k} |d(\sigma) - d_0(\sigma)| < \varepsilon k$, and let us prove $\sum_{\sigma \in \mathbf{N}^{k+1}} |d(\sigma) - d_0(\sigma)| < \varepsilon(k+1)$. We have

$$
\begin{aligned}
&\sum_{\sigma \in \mathbf{N}^{k+1}} |d(\sigma) - d_0(\sigma)| \\
&= \sum_{n \in \mathbf{N}} \sum_{\tau \in \mathbf{N}^k} |d(\tau * n) - d_0(\tau * n)| = \sum_{n \in \mathbf{N}} \sum_{\tau \in \mathbf{N}^k} |d(\tau)d(n) - d_0(\tau)d_0(n)| \\
&= \sum_{n \in \mathbf{N}} \sum_{\tau \in \mathbf{N}^k} |d(\tau)d(n) - d(\tau)d_0(n) + d(\tau)d_0(n) - d_0(\tau)d_0(n)| \\
&\leqslant \sum_{n \in \mathbf{N}} \sum_{\tau \in \mathbf{N}^k} d(\tau)|d(n) - d_0(n)| + \sum_{n \in \mathbf{N}} \sum_{\tau \in \mathbf{N}^k} d_0(n)|d(\tau) - d_0(\tau)| \\
&= \sum_{n \in \mathbf{N}} |d(n) - d_0(n)| \sum_{\tau \in \mathbf{N}^k} d(\tau) + \sum_{n \in \mathbf{N}} d_0(n) \sum_{\tau \in \mathbf{N}^k} |d(\tau) - d_0(\tau)| \\
&= \sum_{n \in \mathbf{N}} |d(n) - d_0(n)| + \sum_{\tau \in \mathbf{N}^k} |d(\tau) - d_0(\tau)| < \varepsilon + \varepsilon k = \varepsilon(k+1).
\end{aligned}
$$

This concludes the proof of Lemma 2.12. □

We return to the proof of the Theorem 2.11. By our assumption there are a polynomial $q(x)$ and a learner Φ such that the following holds: for all $d_0 \in \mathscr{D}$ and for all $m>0$, Φ, on a random sequence σ of length $q(m)$ drawn according the distribution d_0, conjectures, with probability $\alpha > 1-1/m$, (the code of) a distribution d_1 such that $D(d_0,d_1)<1/m$.

We want to prove that for all $m>0$ and for all $d \in \bar{\mathscr{D}}$, the learner Φ, given a sequence of length $q(2m)$ drawn according to the distribution d, conjectures, with probability $\alpha > 1-1/m$, a distribution d_1 such that $D(d,d_1)<1/m$. (Note that the bound $q(x)$ turns into $q(2x)$, but the learning algorithm does not change.)

Let $d_0 \in \mathscr{D}$ such that $D(d,d_0)<1/2m\,q(2m)$. Define

$$A = \left\{ \sigma \in Seq : lth(\sigma) = q(2m) \text{ and } D(h_{\text{fin}}(\Phi(\sigma,2m)),d_0) < \frac{1}{2m} \right\}.$$

From the hypotheses on $\mathscr{D}$, Φ and $q(x)$ we get

$$d_0(A) = \sum_{\sigma \in A} d_0(\sigma) > 1 - \frac{1}{2m}.$$

By Lemma 2.12 we have

$$|d(A) - d_0(A)| \leqslant \sum_{\sigma \in A} |d(\sigma) - d_0(\sigma)| < \frac{q(2m)}{2m\,q(2m)} = \frac{1}{2m},$$

therefore,

$$d(A) > d_0(A) - \frac{1}{2m} > 1 - \frac{1}{2m} - \frac{1}{2m} = 1 - \frac{1}{m}.$$

Moreover, if $\sigma \in A$ and $h_{\text{fin}}(\Phi(\sigma,2m)) = d_1$, then

$$\begin{aligned} D(d,d_1) &= \sum_{x \in \mathbf{N}} |d(x) - d_1(x)| = \sum_{x \in \mathbf{N}} |d(x) - d_0(x) + d_0(x) - d_1(x)| \\ &\leqslant \sum_{x \in \mathbf{N}} |d(x) - d_0(x)| + \sum_{x \in \mathbf{N}} |d_0(x) - d_1(x)| < \frac{1}{2m\,q(2m)} + \frac{1}{2m} \\ &\leqslant \frac{1}{2m} + \frac{1}{2m} = \frac{1}{m}. \end{aligned}$$

So the learner Φ conjectures, with probability $d(A) > 1-1/m$, a distribution d_1 such that $D(d,d_1)<1/m$. □

We introduce another paradigm of PAC learning which heavily depends on representations.

Definition 2.13. Let $\mathscr{D}$ be a countable class of distributions, and let $\mathscr{R} = \langle R, h \rangle$ be a representation of $\mathscr{D}$. We say that $\mathscr{D}$ is PAC learnable wrt the indexes according to the

representation $\mathscr{R}$ of $\mathscr{D}$ iff there are a function $\Phi(\sigma,n): Seq \times \mathbf{N} \xrightarrow{\Phi} R$ and a polynomial $p(x,y)$ such that for all $d \in \mathscr{D}$, if $\bar{r}$ is the shortest representation of d, then, for all $n>0$, and for all $k \geqslant p(l(\bar{r}),n)$, one has

$$d\left(\left\{\sigma \in \mathbf{N}^k : D(h(\Phi(\sigma,n)),d) \geqslant \frac{1}{n}\right\}\right) < \frac{1}{n}.$$

Definition 2.14. $\mathscr{D}$ is computably PAC learnable wrt the indexes according to $\mathscr{R}$ iff the conditions of Definition 2.13 hold, and in addition the learner Φ occurring in Definition 2.13 is P-time computable.

Definition 2.15. $\mathscr{D}$ is (computably) weakly PAC learnable wrt the indexes according to $\mathscr{R}$ iff the conditions of the Definition 2.13 hold with (Φ P-time and) $k \geqslant p(\bar{r}^{\circ},n)$ (where $^{\circ}$ is as in Lemma 2.6) instead of $k \geqslant p(l(\bar{r}),n)$.

Lemma 2.16. *Let $\mathscr{R}=\langle R,h\rangle$, $\mathscr{R}'=\langle R',h'\rangle$ be P-equivalent representations of a class $\mathscr{D}$. Then $\mathscr{D}$ is (computably) PAC learnable wrt the indexes according to $\mathscr{R}$ iff it is such wrt $\mathscr{R}'$.*

Proof. Let f, g be P-time functions such that $h' \circ f_{|R} = h$ and $h \circ g_{|R'} = h'$. Let for every learner Φ, $\Phi^+ = f \circ \Phi$, and $\Phi^- = g \circ \Phi$. Then, if Φ (computably) PAC learns $\mathscr{D}$ wrt the indexes according to $\mathscr{R}$, Φ^+ (computably) PAC learns $\mathscr{D}$ wrt the indexes according to $\mathscr{R}'$, and if Φ (computably) PAC learns $\mathscr{D}$ wrt to the indexes relative to $\mathscr{R}'$, then Φ^- (computably) PAC learns $\mathscr{D}$ wrt to the indexes relative to $\mathscr{R}$. □

Since many natural countable classes of distributions have a natural representation $\mathscr{R}=\langle R,h\rangle$ with R in P, in view of Lemmas 2.6 and 2.16, there is not much loss of generality if we use representations of the form $\langle \mathbf{N},h\rangle$. Whenever we do so, we write d_n for $h(n)$. We also write $\mathscr{D}=\langle d_n : n \in \mathbf{N}\rangle$ to mean that we consider a representation of $\mathscr{D}$ of the form $\langle \mathbf{N},h\rangle$, and that d_n is short for $h(n)$.

3. Non-computable PAC learning

The next result provides for a characterization of non-computable PAC learning.

Theorem 3.1. *Let $\mathscr{D}$ be a class of distributions. The following are equivalent*:

(i) *$\mathscr{D}$ is (non-computably) absolutely PAC learnable.*

(ii) *There are a polynomial $q(k)$ and a family $\{S_d^k : d \in \mathscr{D},\ k \in \mathbf{N}-\{0\}\}$ of subsets of Seq such that*
   (a) *For all $d \in \mathscr{D}$, and for all $k>0$, $S_d^k \subseteq \mathbf{N}^{q(k)}$;*
   (b) *For all $d \in \mathscr{D}$, and for all $k>0$, $d(S_d^k) > 1-1/k$;*
   (c) *If $d,d' \in \mathscr{D}$ and $D(d,d') \geqslant 1/k$, then $S_d^k \cap S_{d'}^k = \emptyset$.*

Proof. (i) $\Rightarrow$ (ii). Let $\Phi(\sigma,k)$ and $p(k)$ be a learner and a polynomial, respectively, such that for all $d\in\mathscr{D}$, and for all $k>0$, if $n\geqslant p(k)$, then

$$(3.1.1)\quad d\left(\left\{\sigma\in\mathbf{N}^n : D(d,h_{\mathrm{fin}}(\Phi(\sigma,k)))\geqslant\frac{1}{k}\right\}\right)<\frac{1}{k}.$$

Let $q(k)=p(2k)$, and let

$$S_d^k=\left\{\sigma\in\mathbf{N}^{q(k)} : D(d,h_{\mathrm{fin}}(\Phi(\sigma,2k)))<\frac{1}{2k}\right\}.$$

Clearly (a) is satisfied.

We verify (b). By (3.1.1), $d(S_d^k)>1-1/2k>1-1/k$.

We verify (c). Argue contrapositively: if $\sigma\in S_d^k\cap S_{d'}^k$ then

$$D(d,d')\leqslant D(d,h_{\mathrm{fin}}(\Phi(\sigma,2k)))+D(d',h_{\mathrm{fin}}(\Phi(\sigma,2k)))<\frac{1}{2k}+\frac{1}{2k}=\frac{1}{k}.$$

(ii) $\Rightarrow$ (i). Let $\{S_d^k : d\in\mathscr{D},\ k\in\mathbf{N}-\{0\}\}$ and $q(k)$ satisfy (a), (b) and (c) in (ii). We define a learner $\Phi(\sigma,k)$ as follows. Fix $\bar{r}\in R_{\mathrm{fin}}$. If $lth(\sigma)<q(2k)$, then $\Phi(\sigma,k)=\bar{r}$. Otherwise, let τ be the sequence constituted of the first $q(2k)$ elements of σ. If $\tau\notin\bigcup_{d\in\mathscr{D}}S_d^{2k}$, then $\Phi(\sigma,k)=\bar{r}$. Otherwise, let $\Phi(\sigma,k)$ be the minimal $r\in R_{\mathrm{fin}}$ for which there is $d'\in\mathscr{D}$ such that $\tau\in S_{d'}^{2k}$, and $D(h_{\mathrm{fin}}(r),d')<1/2k$.

Let $p(k)=q(2k)$. Let σ be a sequence of length $\geqslant p(k)$ drawn at random according to the distribution $d\in\mathscr{D}$. Let τ be the sequence consisting of the first $p(k)$ elements of σ. By condition (b), $\tau\in S_d^{2k}$ with probability $\alpha>1-1/2k$. So, again with probability $>1-1/2k$, $\Phi(\sigma,k)$ outputs an $r\in R_{\mathrm{fin}}$ for which there is $d'\in\mathscr{D}$ such that $\tau\in S_{d'}^{2k}$, and $D(h_{\mathrm{fin}}(r),d')<1/2k$; therefore, $\tau\in S_d^{2k}\cap S_{d'}^{2k}$ and, by condition (c), $D(d,d')<1/2k$. So

$$D(h_{\mathrm{fin}}(\Phi(\sigma,k)),d)\leqslant D(h_{\mathrm{fin}}(\Phi(\sigma,k)),d')+D(d',d)<\frac{1}{2k}+\frac{1}{2k}=\frac{1}{k}.\qquad\square$$

Example 3.2. Let, for $Y\in\mathscr{P}^{<\omega}(\mathbf{N})$, $Y\neq\emptyset$,

$$d_Y(x)=\begin{cases}\dfrac{1}{Card(Y)} & \text{if } x\in Y\\ 0 & \text{otherwise.}\end{cases}$$

The class $\{d_Y : Y\in\mathscr{P}^{<\omega}(\mathbf{N}),\ Y\neq\emptyset\}$ is not absolutely PAC learnable.

Proof. By contradiction we suppose there are a polynomial $q(n)$ and a family $\{S_{d_Y}^n : Y\in\mathscr{P}^{<\omega}(\mathbf{N}),\ Y\neq\emptyset,\ n\in\mathbf{N}-\{0\}\}$ of subsets of Seq as in Theorem 3.1. To simplify notation we write S_Y^n for $S_{d_Y}^n$. We reach a contradiction using the following claims. Let $Y\div Z$ denote the symmetric difference of Y and Z $(Y\backslash Z)\cup(Z\backslash Y)$.

Claim 1. Let $M>1$ be given, and let $Y,Z\subseteq[1,M]$. If $Card(Y\div Z)\geqslant M/n$, then $D(d_Y,d_Z)\geqslant 1/n$.

Proof.

$$D(d_Y, d_Z) \geqslant \sum_{i \in Y-Z} \frac{1}{Card(Y)} + \sum_{i \in Z-Y} \frac{1}{Card(Z)} \geqslant \frac{Card(Y \div Z)}{M} \geqslant \frac{1}{n}.$$

Claim 2. For all $n>0$, and for all $Y \subseteq [1,M]$, there are at most $2^{M/2}$ sets $Z \subseteq [1,M]$ such that $D(d_Y, d_Z) < 1/n$.

Proof. Any such Z is uniquely determined by Y and $Y \div Z$. Now Y is fixed, and by Claim 1, the number of possible $Y \div Z$ is bounded by the number of subsets of $[1,M]$ of cardinality $\leqslant M/n$. For each i there are $\binom{M}{i}$ subsets of $[1,M]$ of cardinality i. So there are $\sum_{i \leqslant \frac{M}{n}} \binom{M}{i}$ subsets of $[1,M]$ of cardinality $\leqslant M/n$. Note that

$$\left(\frac{1}{n}\right)^{M/n} \sum_{i \leqslant \frac{M}{n}} \binom{M}{i} \leqslant \sum_{i \leqslant \frac{M}{n}} \left(\frac{1}{n}\right)^{i} \binom{M}{i} \leqslant \sum_{i \leqslant M} \binom{M}{i} \left(\frac{1}{n}\right)^{i}$$

$$= \left(1 + \frac{1}{n}\right)^{M} \leqslant \exp\left(\frac{M}{n}\right).$$

Hence, $\sum_{i \leqslant \frac{M}{n}} \binom{M}{i} \leqslant n^{M/n} \exp(\frac{M}{n}) = 2^{M/n(\log(e)+\log(n))}$.

Then, for n sufficiently large, we have

$$\sum_{i \leqslant \frac{M}{n}} \binom{M}{i} \leqslant 2^{M/2}.$$

Claim 3. There are at least $H = 2^{M/2}$ subsets $Y_1 \cdots Y_H$ of $[1,M]$ such that for $i \neq j$, $D(d_{Y_i}, d_{Y_j}) \geqslant 1/n$.

Proof. There are 2^M subsets of $[1,M]$, and for each such subset Y there are at most $2^{M/2}$ sets $Z \subseteq [1,M]$ such that $D(d_Y, d_Z) < 1/n$. Hence, there are at least $H = 2^M/2^{M/2} = 2^{M/2}$ subsets $Y_1 \cdots Y_H$ such that for $i \neq j$, $D(d_{Y_i}, d_{Y_j}) \geqslant 1/n$. This concludes the proof of Claim 3.

We conclude the proof of Example 3.2.

We can find M such that $H = 2^{M/2} > M^{q(n)}$ (let e.g. $M > (2\,q(n))^2$). Let $Y_1 \cdots Y_H$ be as in Claim 3. For $i, j \leqslant H$ one has

$d_{Y_i}(S^n_{Y_i}) > 1 - \frac{1}{n}$, and, if $i \neq j$, then $S^n_{Y_i} \cap S^n_{Y_j} = \emptyset$.

Now if $z \notin [1,M]$, then for $i \leqslant H$ one has $d_{Y_i}(z) = 0$. Thus, for $i = 1 \cdots H$, $S^n_{Y_i} \cap [1,M]^{q(n)} \neq \emptyset$.

Since, for $i \neq j$ $S^n_{Y_i} \cap S^n_{Y_j} = \emptyset$, we get $M^{q(n)} = Card[1,M]^{q(n)} \geqslant H$, a contradiction. □

Theorem 3.3. *Let $\mathscr{D} = \langle d_i : i \in \mathbf{N} \rangle$ be any countable class of distributions. Then $\mathscr{D}$ is (non-computably) PAC learnable wrt the indexes.*

Proof. Let $n>0$ be given. For each i,j such that $D(d_i,d_j)\geqslant 1/n$, let $C_{ij}=\{x: d_i(x)>d_j(x)\}$. (Thus, $C_{ji}=\{x: d_j(x)>d_i(x)\}$.)
Note that $d_i(C_{ij})-d_j(C_{ij})=d_j(C_{ji})-d_i(C_{ji})\geqslant 1/2n$.
Let $\sigma\in Seq$ and $n>0$ be given. Let

$$A_{\sigma,n}=\left\{i\leqslant 2^{lth(\sigma)/32n^2}:\forall j<i\left(D(d_i,d_j)\geqslant\frac{1}{n}\right)\right.$$
$$\left.\Rightarrow\left(d_i(C_{ij})-f_\sigma(C_{ij})\leqslant\frac{1}{4n}\right)\right\}.$$

Define

$$\Phi(\sigma,n)=\begin{cases}\max(A_{\sigma,n}) & \text{if } A_{\sigma,n}\neq\emptyset,\\ 0 & \text{otherwise.}\end{cases}$$

(Note that we do not claim that Φ is computable.)

We claim that Φ PAC learns $\mathscr{D}$ wrt the indexes. Let σ be drawn at random according to the distribution d_i, and let $m=lth(\sigma)$. It is sufficient to prove that if $m\geqslant 32n^2(l(i)+\ln(n))$, then, with probability $\geqslant 1-1/n$, Φ conjectures an index j such that $D(d_i,d_j)<1/n$. By definition, $\Phi(\sigma,n)$ can only output an index j such that $0\leqslant j\leqslant 2^{m/32n^2}$. So, if $m\geqslant(32n^2)l(i)$, then i is a possible output of Φ. Now let E_0, E_1 be the events defined as follows:
E_0: There is $j<i$ such that $D(d_i,d_j)\geqslant 1/n$ and $d_i(C_{ij})-f_\sigma(C_{ij})>1/4n$.
E_1: There is j such that $i<j<2^{m/32n^2}$, $D(d_i,d_j)\geqslant 1/n$, and for all

$$h<j,\quad\text{if } D(d_h,d_j)\geqslant\frac{1}{n},\quad\text{then } d_j(C_{jh})-f_\sigma(C_{jh})\leqslant\frac{1}{4n}.$$

If none of E_0, E_1 occurs, then $\Phi(\sigma,n)$ outputs a j such that $D(d_i,d_j)<1/n$. Thus, we have to upper bound the probability of $E_0\cup E_1$ in order to evaluate the probability that Φ's error is $\geqslant 1/n$.

By Chernoff's bound, for each single $j<i$ one has

$$Pr\left(d_i(C_{ij})-f_\sigma(C_{ij})>\frac{1}{4n}\right)<\exp\left(-\frac{m}{16n^2}\right).$$

Thus $Pr(E_0)\leqslant i\exp(-m/16n^2)$.

Now for each j such that $i<j<2^{m/32n^2}$ and $D(d_i,d_j)\geqslant 1/n$ consider the event E'_j: For all $h<j$, if $D(d_h,d_j)\geqslant 1/n$, then $d_j(C_{jh})-f_\sigma(C_{jh})\leqslant 1/4n$.

Clearly, E'_j implies $d_j(C_{ji})-f_\sigma(C_{ji})\leqslant 1/4n$. Since $d_j(C_{ji})-d_i(C_{ji})\geqslant 1/2n$, E'_j implies $f_\sigma(C_{ji})-d_i(C_{ji})\geqslant 1/4n$. By Chernoff's bound, this occurs with probability $\leqslant\exp(-m/16n^2)$. Since E_1 is the union of all possible E'_j, E_1 occurs with probability $\leqslant(2^{m/32n^2}-i)\exp(-m/16n^2)$. Thus

$$Pr(E_0\cup E_1)\leqslant(2^{m/32n^2}-i+i)\exp\left(-\frac{m}{16n^2}\right)=2^{m/32n^2}\exp\left(-\frac{m}{16n^2}\right).$$

Thus

$$Pr(E_0 \cup E_1) \leqslant \exp\left(\frac{m}{32n^2} - \frac{m}{16n^2}\right) = \exp\left(-\frac{m}{32n^2}\right).$$

It follows that if $m \geqslant 32n^2 \ln(n)$, and $m \geqslant 32n^2 l(i)$, then with probability $> 1 - 1/n$ Φ conjectures a j such that $D(d_i, d_j) < 1/n$. So, for all n, i, $32n^2(l(i) + ln(n))$ examples suffice to identify d_i with probability $> 1 - 1/n$ and with error $< 1/n$. □

4. Absolute computable PAC learning: a frequentist approach

In the most natural examples of absolutely PAC learnable classes of distributions we met, the best strategy seems to be the frequentist one, which consists in guessing, on every sequence σ of examples, the finite distribution f_σ. (There are examples of absolutely computably PAC learnable classes of distributions which cannot be learned by means of this strategy, but these classes are very *ad hoc* ones.)

Definition 4.1. The frequentist learner is the learner Φ_f defined, for every $\sigma \in Seq$ and for every $k \in \mathbf{N} - \{0\}$, by $\Phi_f(\sigma, k) = (h_{\text{fin}})^{-1}(f_\sigma)$ (cf. Definition 2.8 for the definition of h_{fin}).

Clearly, the frequentist learner works in P-time.

Definition 4.2. A class $\mathscr{D}$ of distributions is said to be polynomially localized iff there exists a polynomial $p(x)$ such that, for all $d \in \mathscr{D}$ and for all $n \in \mathbf{N} - \{0\}$, there exists $A_{d,n} \subseteq \mathbf{N}$ with $Card(A_{d,n}) \leqslant p(n)$, and $\sum_{x \in A_{d,n}} d(x) > 1 - 1/n$.

Theorem 4.3. *Let $\mathscr{D}$ be a class of distributions. The following are equivalent*:
(i) *$\mathscr{D}$ is computably absolute PAC learnable by the frequentist learner.*
(ii) *$\mathscr{D}$ is polynomially localized.*

Proof. (ii) $\Rightarrow$ (i). Let $d \in \mathscr{D}$, $n \in \mathbf{N} - \{0\}$, and $A_{d,n} \subseteq \mathbf{N}$ such that $Card(A_{d,n}) \leqslant p(8n)$, and $\sum_{x \in A_{d,n}} d(x) > 1 - 1/8n$.

Applying the Chernoff's bound to the event "$x \in A_{d,n}$", we get

$$Pr\left(\left|\sum_{x \in A_{d,n}} f_\sigma(x) - \sum_{x \in A_{d,n}} d(x)\right| \geqslant \frac{1}{8n}\right) \leqslant 2\exp\left(-\frac{1}{64n^2} lth(\sigma)\right).$$

Hence

$$Pr\left(\sum_{x \in A_{d,n}} f_\sigma(x) \leqslant 1 - \frac{1}{4n}\right) \leqslant 2\exp\left(-\frac{1}{64n^2} lth(\sigma)\right)$$

which implies

$$Pr\left(\sum_{x\notin A_{d,n}} f_\sigma(x) \geqslant \frac{1}{4n}\right) \leqslant 2\exp\left(-\frac{1}{64n^2} lth(\sigma)\right).$$

Applying Chernoff's bound, for all $x\in A_{d,n}$, we get

$$Pr\left(|f_\sigma(x)-d(x)| \geqslant \frac{1}{4np(8n)}\right) \leqslant 2\exp\left(-\frac{lth(\sigma)}{16n^2(p(8n))^2}\right).$$

Let $q(n)=64n^2+16n^2(p(8n))^2$ and let E be the event

$$\text{either } \sum_{x\notin A_{d,n}} f_\sigma(x) \geqslant \frac{1}{4n} \quad \text{or} \quad \exists x\in A_{d,n} : |f_\sigma(x)-d(x)| \geqslant \frac{1}{4np(8n)}.$$

We get

$$\begin{aligned} Pr(E) &\leqslant 2\exp\left(-\frac{lth(\sigma)}{64n^2}\right) + p(8n)2\exp\left(-\frac{lth(\sigma)}{16n^2(p(8n))^2}\right) \\ &\leqslant (2+2p(8n))\exp\left(-\frac{lth(\sigma)}{q(n)}\right). \end{aligned}$$

If $lth(\sigma)\geqslant q(n)(\ln(n)+\ln(2+2p(8n)))$, then $Pr(E)<1/n$.

On the other hand, if E does not occur, we have

$$\begin{aligned} \sum_{x\in\mathbf{N}} |f_\sigma(x)-d(x)| &= \sum_{x\in A_{d,n}} |f_\sigma(x)-d(x)| + \sum_{x\notin A_{d,n}} |f_\sigma(x)-d(x)| \\ &< p(8n)\frac{1}{4np(8n)} + \sum_{x\notin A_{d,n}} f_\sigma(x) + \sum_{x\notin A_{d,n}} d(x) < \frac{1}{4n}+\frac{1}{4n}+\frac{1}{8n} < \frac{1}{n}. \end{aligned}$$

So $D(d,f_\sigma)<1/n$. Thus, if $lth(\sigma)\geqslant q(n)(\ln(n)+\ln(2+2p(8n)))$, then

$$Pr\left(D(d,f_\sigma) \geqslant \frac{1}{n}\right) < \frac{1}{n}.$$

(i) $\Rightarrow$ (ii). Suppose that $\mathscr{D}$ is not polynomially localized. Let, by contradiction, $g(n)$ be a polynomial such that for all $d\in\mathscr{D}$, the probability of drawing at random wrt the distribution d a sequence σ of length at least $g(n)$ such that $D(d,f_\sigma)\geqslant 1/n$ is less than $1/n$.

Since $\mathscr{D}$ is not polynomially localized, there exist $d\in\mathscr{D}$ and $\bar{n}\in\mathbf{N}$ such that, for all $A\subseteq\mathbf{N}$ of cardinality at most $g(\bar{n})$, one has

$$\sum_{x\in A} d(x) \leqslant 1-\frac{1}{\bar{n}}.$$

Let $\sigma\in Seq$ such that $lth(\sigma)=g(\bar{n})$, and let $A=\{x : f_\sigma(x)\neq 0\}$.

Clearly, $Card(A) \leqslant g(\bar{n})$, therefore, $\sum_{x \notin A} d(x) \geqslant 1/\bar{n}$.

Since $\sum_{x \notin A} f_\sigma(x) = 0$, we get $D(f_\sigma, d) \geqslant 1/\bar{n}$.

As this result does not depend on the choice of $\sigma \in \mathbf{N}^{g(\bar{n})}$, the frequentist learner does not PAC learn $\mathscr{D}$, a contradiction. □

Example 4.4. Let, for every non-empty subset X of $\mathbf{N}$, $C_X = \sum_{n \in X} \frac{1}{2^{n+1}}$, and let, for all $k \in \mathbf{N}$:

$$d_X(k) = \begin{cases} \dfrac{1}{C_X 2^{k+1}} & \text{if } k \in X, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the class $\mathscr{D} = \{d_X : X \subseteq \mathbf{N},\ X \neq \emptyset\}$ is an absolutely computably PAC learnable class of distributions.

Proof. Let $X \subseteq \mathbf{N}$, $X \neq \emptyset$ be given. Let $\bar{i}$ be the minimal element of X. Then $C_X \geqslant 1/2^{\bar{i}+1}$. Hence, for all $i \in \mathbf{N}$, $d_X(i) \leqslant 2^{\bar{i}}/2^i$.

Now let $m = \bar{i} + n$. One has

$$\sum_{i > m} d_X(i) \leqslant \frac{2^{\bar{i}}}{2^m} = \frac{1}{2^n} < \frac{1}{n}.$$

Thus, $\sum_{\bar{i} \leqslant i \leqslant \bar{i}+n} d_X(i) > 1 - 1/n$. It follows that, for all $X \subseteq \mathbf{N}$, $X \neq \emptyset$, and for all $n \in \mathbf{N} - \{0\}$, there is a set $A_{X,n}$ of $\leqslant n$ elements such that $\sum_{i \in A_{X,n}} d_X(i) > 1 - 1/n$.

Hence, $\mathscr{D}$ is polynomially localized, therefore, by Theorem 4.3, it is absolutely computably PAC learnable. □

Remark. Note that $\mathscr{D} = \{d_X : X \neq \emptyset,\ X \subseteq \mathbf{N}\}$ has the following properties:

(i) $\mathscr{D}$ is uncountable;

(ii) for all $X \neq \emptyset$, $X \subseteq \mathbf{N}$, there is $d_X \in \mathscr{D}$ such that $X = \{n : d_X(n) \neq 0\}$.

5. Computable PAC learning wrt the indexes

In this section, we prove that also in the computable case PAC learning wrt the indexes is quite different from absolute PAC learning. We start with a result which is in sharp contrast with Example 3.2.

Theorem 5.1. *The class $\mathscr{D}_{\text{fin}}$ of finite distributions is computably PAC learnable wrt the indexes according to the canonical representation* (*defined in Definition* 2.8).

Proof. We claim that the frequentist learner Φ_f PAC learns $\mathscr{D}_{\text{fin}}$ wrt the indexes.

Let $n > 0$ be given, let σ be drawn at random according to any distribution $d \in \mathscr{D}_{\text{fin}}$, let $m = lth(\sigma)$, and let r be the code of d. Let $S_d = \{x : d(x) > 0\}$. By Chernoff's bounds,

for each $x \in S_d$ we have

$$Pr\left(|d(x) - f_\sigma(x)| \geqslant \frac{1}{nl(r)}\right) \leqslant 2\exp\left(-\frac{m}{n^2(l(r))^2}\right).$$

Now it follows from the definition of $\mathscr{R}_{\text{fin}}$ that $Card(S_d) \leqslant l(r)$. Thus the probability that there is an $x \in S_d$ such that $|d(x) - f_\sigma(x)| \geqslant 1/nl(r)$ is at most $2l(r)\exp(-m/n^2(l(r))^2)$.

If $m \geqslant n^2 l(r)^2(\ln(l(r)) + \ln(n) + \ln(3))$, then with probability $\geqslant 1 - 2l(r)\exp(-(\ln(l(r)) + \ln(n) + \ln(3))) = 1 - 2/3n > 1 - 1/n$ we have that for all $x \in S_d$, $|d(x) - f_\sigma(x)| < 1/nl(r)$.

Now if $d(x) = 0$, then with probability 1 one has $f_\sigma(x) = 0$.

It follows that with probability $> 1 - 1/n$ one has

$$D(d, f_\sigma) = \sum_{x \in S_d} |d(x) - f_\sigma(x)| < \frac{Card(S_d)}{nl(r)} < \frac{1}{n}.$$

This completes the proof. □

In general, the frequentist learner cannot learn wrt the indexes, because it is possible that he is not allowed to guess any finite distribution. A frequentist approach to learning wrt the indexes would be the following: given a sequence σ of examples, the learner tries to make a guess from the space of hypotheses which is as close as possible to f_σ. However, in general finding such a guess is not feasible in P-time.

In this section, we investigate such strategy of learning, and we connect it with the $NP = RP$ problem. We start from a positive result concerning weak PAC learning.

Theorem 5.2. *Any polynomially localized class $\mathscr{D}$ of distributions having a P-representation is computably weakly PAC learnable wrt the indexes.*

Proof. Let $\mathscr{D} = \langle d_i : i \in \mathbf{N}\rangle$ be as in the statement of the present theorem, and let $h(i,x,n)$ be a P-time computable function such that for all $i,x,n \in \mathbf{N}$, $n > 0$, $|h(i,x,n) - d_i(x)| < 1/n$. Let $p(n)$ be a polynomial such that for all $i,n \in \mathbf{N}$, $n > 0$, there is $A_{i,n} \subseteq \mathbf{N}$ such that

$$Card(A_{i,n}) \leqslant p(n), \quad \text{and} \quad \sum_{x \in A_{i,n}} d_i(x) > 1 - \frac{1}{n}.$$

We can suppose wlog that $p(n) \geqslant 2n$. Let $d_i \in \mathscr{D}$ be given. We know (Theorem 4.3) that there is a polynomial $q(x)$ such that for all $k \geqslant q(n)$,

$$d_i\left(\left\{\sigma \in Seq : lth(\sigma) = k \text{ and } D(d_i, f_\sigma) \geqslant \frac{1}{n}\right\}\right) < \frac{1}{n}.$$

Now let $r(n) = (p(6n))^2$, $s(n) = q(r(n))$. If $k \geqslant s(n)$, then with probability $> 1 - 1/r(n)$, whenever a σ of length k is chosen according to the distribution d_i, one has $D(d_i, f_\sigma) < 1/r(n)$.

We show that this implies that whenever d_j is such that for all $x \in \mathrm{range}(\sigma)$, $|f_\sigma(x) - d_j(x)| < 1/r(n)$, $D(d_i, d_j) < 1/n$. First of all, with probability $> 1 - 1/r(n)$, one has

$$f_\sigma(A_{i,6n}) > d_i(A_{i,6n}) - \frac{1}{r(n)} > 1 - \frac{1}{6n} - \frac{1}{r(n)} \geqslant 1 - \frac{1}{4n}.$$

Let $range(\sigma) = \{x \in \mathbf{N} : f_\sigma(x) \neq 0\}$. Let $A_{i,n}(\sigma) = A_{i,6n} \cap range(\sigma)$. Clearly, $f_\sigma(A_{i,n}(\sigma)) = f_\sigma(A_{i,6n}) > 1 - 1/4n$. Again with probability $> 1 - 1/r(n)$, we obtain that, if $d_j f \in \mathscr{D}$ is such that for all $x \in range(\sigma)$, $|f_\sigma(x) - d_j(x)| < 1/r(n)$, then

$$(5.2.1) \quad \sum_{x \in A_{i,n}(\sigma)} |f_\sigma(x) - d_j(x)| < \frac{1}{p(6n)}.$$

$$(5.2.2) \quad d_j(A_{i,n}(\sigma)) > f_\sigma(A_{i,n}(\sigma)) - \frac{1}{p(6n)} > 1 - \frac{1}{4n} - \frac{1}{p(6n)} \geqslant 1 - \frac{1}{3n}, \quad \text{and}$$

$$d_j(\mathbf{N} - A_{i,n}(\sigma)) < \frac{1}{3n}.$$

$$(5.2.3) \quad D(d_j, f_\sigma) \leqslant \sum_{x \in A_{i,n}(\sigma)} |f_\sigma(x) - d_j(x)| + d_j(\mathbf{N} - A_{i,n}(\sigma)) + f_\sigma(\mathbf{N} - A_{i,n}(\sigma)) < \frac{1}{p(6n)} + \frac{1}{3n} + \frac{1}{4n} \leqslant \frac{2}{3n}.$$

$$(5.2.4) \quad D(d_i, d_j) \leqslant D(d_i, f_\sigma) + D(f_\sigma, d_j) < \frac{1}{n}.$$

Summing up, we have shown the following:

Fact (5.2.5). If we choose a sequence σ of length $\geqslant s(n)$ according to the distribution d_i, then, with probability $> 1 - 1/r(n)$, $D(f_\sigma, d_i) < 1/r(n)$, and, whenever d_j is such that for all $x \in range(\sigma)$, $|f_\sigma(x) - d_j(x)| < 1/r(n)$, one has $D(d_i, d_j) < 1/n$.

Now let

$$B_{\sigma,n} = \left\{ j < lth(\sigma) : \forall x \in range(\sigma) | f_\sigma(x) - h(j, x, 8r(n))| < \frac{1}{2r(n)} \right\}.$$

Define the following learner

$$\Phi(\sigma, n) = \begin{cases} \min(B_{\sigma,n}) & \text{if } B_{\sigma,n} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, Φ is P-time. Moreover, once $B_{\sigma,n} \neq \emptyset$, $\Phi(\sigma, n)$ outputs a $j \in B_{\sigma,n}$. Therefore for all $x \in range(\sigma)$ we have

$$(5.2.6) \quad |f_\sigma(x) - d_j(x)| \leqslant |f_\sigma(x) - h(j, x, 8r(n))| + |d_j(x) - h(j, x, 8r(n))| < \frac{1}{2r(n)} + \frac{1}{8r(n)} < \frac{1}{r(n)}.$$

By Fact (5.2.5), if σ is chosen at random according to the distribution d_i, and if $lth(\sigma) \geqslant s(n)$, then, with probability $> 1 - 1/r(n)$ we have $D(f_\sigma, d_i) < 1/r(n)$, and so we have

$$(5.2.7) \quad D(d_{\Phi(\sigma,n)}, d_i) < \frac{1}{n}.$$

Now we show that $D(f_\sigma, d_i) < 1/4r(n)$ implies that $B_{\sigma,n} \neq \emptyset$, with probability $> 1 - 1/4r(n)$. Let $t(i,n) = i + q(4r(n))$. Clearly, $t(i,n) > s(n)$. If we choose a sequence σ of length $\geqslant t(i,n)$ at random according to the distribution d_i, then with probability $> 1 - 1/4r(n)$ we obtain $D(f_\sigma, d_i) < 1/4r(n)$. Hence, for all $x \in range(\sigma)$,

$$\begin{aligned} |f_\sigma(x) - h(i,x,8r(n))| &\leqslant |f_\sigma(x) - d_i(x)| + |d_i(x) - h(i,x,8r(n))| \\ &< \frac{1}{4r(n)} + \frac{1}{8r(n)} < \frac{1}{2r(n)}. \end{aligned}$$

Since $i < lth(\sigma)$, with probability $> 1 - 1/4r(n)$, $i \in B_{\sigma,n}$, and $B_{\sigma,n} \neq \emptyset$. So with probability $> 1 - 1/4r(n)$, $\Phi(\sigma,n)$ outputs a $j \in B_{\sigma,n}$, and by (5.2.7) with probability $> 1 - 1/4r(n) - 1/r(n) > 1 - 1/n$, such a j also satisfies $D(d_i, d_j) < 1/n$. So $\mathscr{D}$ is computably weakly PAC learnable wrt the indexes. □

We would like to extend Theorem 5.2 to computable learning wrt the indexes. However, we will prove that this claim is *nearly* equivalent to $NP = RP$. This is contrast with the case of absolute PAC learning: being polynomially localized is a sufficient condition for absolute computable PAC learnability in Theorem 4.3.

Definition 5.3. Let $\mathscr{D}$, $h(i,x,n)$, $r(n)$, etc. be as in the proof of Theorem 5.2. We define, for $\sigma \in Seq$ and for $n \in \mathbf{N} - \{0\}$,

$$C_{\sigma,n} = \left\{ i \leqslant 2^{lth(\sigma)} : \forall x \in range(\sigma) |f_\sigma(x) - h(i,x,8r(n))| < \frac{1}{2r(n)} \right\}$$

$$\Psi(\sigma,n) = \begin{cases} \min(C_{\sigma,n}) & \text{if } C_{\sigma,n} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Notice that, due to the presence of a not logarithmically bounded μ operator in its definition, Ψ might fail to be in P. ($2^{lth(\sigma)}$ might have more or less the same size as σ.)

Lemma 5.4. *Let Ψ as in Definition* 5.3, *and let $\mathscr{D}$, $h(i,x,n)$, $r(n)$ etc. be as in Theorem* 5.2.

For all $n, i \in \mathbf{N}$, $n > 0$, if σ is a sequence of $q(4r(n) + l(i))$ examples chosen according to the distribution d_i, then with probability $> 1 - 1/n$, one has $D(d_{\Psi(\sigma,n)}, d_i) < 1/n$.

Thus, if Ψ is P-time, then $\mathscr{D}$ is computably PAC learnable wrt the indexes.

Proof. If $lth(\sigma) \geqslant l(i)+1$, then $2^{lth(\sigma)} > 2^{l(i)} \geqslant i$. Hence i is a possible output of the function Ψ introduced in Definition 5.3. Thus, imitating the proof of Theorem 5.2 with Ψ in place of Φ and $C_{\sigma,n}$ in place of $B_{\sigma,n}$, we see that if σ is chosen at random according to the distribution d_i and $lth(\sigma) \geqslant q(4r(n)+l(i))$, then with probability $>1-1/n$ we have $D(d_{\Psi(\sigma,n)}, d_i) < 1/n$. □

Corollary 5.5. *If $P=NP$, then the learner Ψ introduced in Definition* 5.3 *is P-time computable*; *therefore*, *every polynomially localized class of distributions $\mathscr{D}$ having a P-representation is computably PAC learnable wrt the indexes.*

Proof. Let $O(a,b,n,\sigma)$ be the oracle defined by

$$O(a,b,n,\sigma) \equiv \exists i \in [a,b]\left(\forall x \in range(\sigma)\ |f_\sigma(x) - h(i,x,8r(n))| < \frac{1}{2r(n)}\right),$$

where a,b range over $[0,2^{lth(\sigma)}]$. Then, one can compute Ψ by a binary search using $lth(\sigma)$ calls to the oracle $O(a,b,n,\sigma)$. Since $O(a,b,n,\sigma)$ is in NP, if $P=NP$, then Ψ is P-time. □

Lemma 5.6. *Suppose that $NP=RP$. Then the learner Ψ introduced in Definition* 5.3 *can be computed in polynomial time by a probabilistic recursive function. In other words*, *there are a probabilistic recursive function $\chi(\sigma,n,\tau) : Seq \times (\mathbf{N}-\{0\}) \times Bseq \xrightarrow{\chi} \mathbf{N}$ which works in time polynomial in n, $l(\sigma^\sharp)$, $l(\tau)$ and a polynomial $Q(x,y)$ such that, whenever $lth(\tau) \geqslant Q(lth(\sigma),n)$, with probability $>1-1/n$, one has $\chi(\sigma,n,\tau) = \Psi(\sigma,n)$.*

Proof. Let $\Gamma(a,b,n,\sigma,\tau) : \mathbf{N}^2 \times (\mathbf{N}-\{0\}) \times Seq \times Bseq \xrightarrow{\Gamma} \{0,1\}$ be a probabilistic recursive function which works in P-time in a, b, n, $l(\sigma^\sharp)$, $lth(\tau)$, and let $R(a,b,n,x)$ be a polynomial such that, whenever $lth(\tau) \geqslant R(a,b,n,lth(\sigma))$, with probability $>1-1/2n\,lth(\sigma)$, the function $\Gamma(a,b,n,\sigma,\tau)$ gives the same answer as $O(a,b,n,\sigma)$.

We can simulate the computation of $\Psi(\sigma,n)$ replacing each call to the oracle $O(a,b,n,\sigma)$ ($lth(\sigma)$ calls in total) by a computation of $\Gamma(a,b,n,\sigma,\tau)$, where τ is a binary sequence of length at least $R(a,b,n,lth(\sigma))$, chosen with $lth(\tau)$ random draws by means of a fair coin. Note that $l(a)$ and $l(b)$ are $\leqslant l(\sigma^\sharp)$. So, the whole computation is P-time in n and $l(\sigma^\sharp)$.

Let α be the probability that Γ gives the same answer to each call as O (i.e. that Γ always gives the correct answer). By Proposition 2.1 we have

$$\alpha > \left(1-\frac{1}{2n\,lth(\sigma)}\right)^{lth(\sigma)} \geqslant \left(1-\frac{1}{2n\,lth(\sigma)}\right)^{(-2n\,lth(\sigma))(-1/2n)}$$

$$\geqslant \exp\left(-\frac{1}{n}\right) \geqslant 1-\frac{1}{n}.$$

Now, consider the algorithm $\chi(\sigma,n,\tau)$ that simulates the behavior of $\Psi(\sigma,2n)$ by replacing each call to the oracle $O(a,b,2n,\sigma)$ by a computation of $\Gamma(a,b,2n,\sigma,\tau)$.

Again, the computation is polynomial in $l(\sigma^{\sharp})$ and n. Moreover, with probability $>1-1/2n$, $\chi(\sigma,n,\tau)$ simulates the function $\Psi(\sigma,2n)$. Let $r(n)$ and $q(x)$ be as in Lemma 5.4. If σ is chosen according to the distribution d_i, and if $lth(\sigma)\geqslant q(4r(2n)+l(i))$, then, by Lemma 5.4, with probability $>1-1/2n$, one has $D(d_i,d_{\Psi(\sigma,2n)})<1/2n$.

Thus the probability that either our algorithm fails to simulate $\Psi(\sigma,2n)$, or $D(d_i,\, d_{\Psi(\sigma,2n)})\geqslant 1/2n$ is less than $1/2n+1/2n=1/n$.

So, with probability $>1-1/n$, our algorithm computes a j s.t. $D(d_i,d_j)<1/2n$. □

As an immediate consequence we obtain

Theorem 5.7. *If $NP=RP$, then every polynomially localized P-representable class of distributions is computably PAC learnable wrt the indexes by a probabilistic recursive function. In other words, there are a probabilistic recursive function $\chi(\sigma,\tau,n)$ and polynomials $S(x,n),T(y,n)$ such that, whenever σ is chosen according to the distribution d_i, $lth(\sigma)\geqslant S(l(i),n)$, and τ is a binary sequence of length $\geqslant T(lth(\sigma),n)$ chosen at random using a fair coin, with probability $>1-1/n$, $\chi(\sigma,\tau,n)$ outputs a j such that $D(d_i,d_j)<1/n$.*

One might ask whether one can strengthen Theorem 5.7, proving that if $NP=RP$, then any polynomially localized P represented class of distributions is computably PAC learnable wrt to the indexes by a usual recursive function (i.e., by one which works without coins). We prove that this is true under a rather natural additional assumption.

Definition 5.8. A class $\mathscr{D}$ of distributions is said to be atomless iff there is a real number $\alpha<1$ such that for all $d\in\mathscr{D}$ and for all $n\in\mathbf{N}$ one has $d(n)<\alpha$.

Theorem 5.9. *Let $\mathscr{D}$ be as in Theorem* 5.7, *and suppose that in addition $\mathscr{D}$ is atomless. If $NP=RP$, then $\mathscr{D}$ is computably PAC learnable wrt the indexes* (*by means of a usual recursive function*).

Proof. The idea is that we can simulate with high probability a sufficiently long sequence of coin flips by means of a sequence of random examples which is only polynomially longer. We proceed as follows. We ask examples in pairs: $a_0,a_1;\cdots;a_{2n},a_{2n+1}$ until we get $a_{2n}\neq a_{2n+1}$. As soon as we get such a pair $a_{2n}\neq a_{2n+1}$, if $a_{2n}<a_{2n+1}$ we add a 0 to the coin sequence; if $a_{2n}>a_{2n+1}$ we add a 1. Since the probability of drawing a number a_{2n} followed by a different number a_{2n+1} equals the probability of drawing a_{2n+1} followed by a_{2n}, 0 and 1 have the same probability. Moreover, for every n, the probability of the existence of an $i\leqslant n$ such that $a_{2i}\neq a_{2i+1}$ is $\geqslant 1-\alpha^n$. Thus, given natural numbers h,k, the probability that for k times one gets one bit upon seeing at most $2h$ examples is $\geqslant(1-\alpha^h)^k$. If $h\geqslant \ln(2)/\ln(1/\alpha)$, then $\alpha^h\leqslant 1/2$, and by Proposition 2.1 $(1-\alpha^h)^k\geqslant\exp(-2k\alpha^h)$. If we need such a probability to be $\geqslant 1-1/m$, it is sufficient that $\exp(-2k\alpha^h)\geqslant\exp(-1/m)$, i.e., that $h\geqslant\ln(2mk)/\ln(1/\alpha)$. Thus, in order to obtain a coin sequence of length at least k with probability $\geqslant 1-1/m$,

it is sufficient to ask r examples, where r is given by

$$(5.9.1) \quad r = 2k\frac{\ln(2mk)}{\ln(1/\alpha)}.$$

Now for any $n, i \in \mathbf{N}$, the length k of the coin sequence needed by the learner in Theorem 5.7 in order to identify a distribution $d_i \in \mathscr{D}$ with an error $< 1/2n$ and with probability $> 1 - 1/2n$ is polynomial in n and in $l(i)$, say, $k = P(n, l(i))$. Letting, in (5.9.1), $m = 2n$ and $k = P(n, l(i))$, we obtain that after

$$Q(n, l(i)) = 2P(n, l(i))\frac{\ln(4nP(n, l(i)))}{\ln(1/\alpha)}$$

examples, with probability $\geqslant 1 - 1/2n$ a coin sequence of the desired length $P(n, l(i))$ is produced. But in this case the learner identifies d_i with an error $< 1/2n$ with probability $> 1 - 1/2n$. So, the probability that either a coin sequence of the desired length is not produced, or the learner fails to identify d_i with an error $< 1/2n$ is $< 1/n$. □

We are going to prove the converse of Theorem 5.7 that is

Theorem 5.10. *If $NP \neq RP$, then there is a P-representable polynomially localized atomless class of distributions, which is not computably PAC learnable wrt the indexes, not even by a probabilistic P-time recursive function.*

Proof. Let $P(x) \equiv \exists y \leqslant x\, R(x, y)$, be any NP complete predicate, where R is in P. Consider a P-time computable bijection $\bullet$ from $\mathbf{N}^2$ onto $\mathbf{N}$ whose inverse $(\pi_1(x), \pi_2(x))$ is in turn in P. Also, assume that $\bullet$ is increasing in both arguments.

Let d_i be defined as follows

$$d_i(x) = \begin{cases} \frac{1}{2} & \text{if } x = \pi_1(i), \\ \frac{1}{2} & \text{if } x = \pi_1(i) + 1, \pi_2(i) \leqslant \pi_1(i) \text{ and } R(\pi_1(i), \pi_2(i)), \\ \frac{1}{2} & \text{if } x = \pi_1(i) + 2, \pi_2(i) \leqslant \pi_1(i) \text{ and } \neg R(\pi_1(i), \pi_2(i)), \\ \frac{1}{2} & \text{if } x = \pi_1(i) + 3, \text{ and } \pi_2(i) > \pi_1(i), \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathscr{D} = \langle d_i : i \in \mathbf{N} \rangle$ and $h(i) = d_i$. Clearly, $\langle \mathbf{N}, h \rangle$ is a P representation of $\mathscr{D}$. Moreover $\mathscr{D}$ is polynomially localized and atomless. Note that for all $i \in \mathbf{N}$, $P(i)$ is true iff there exists such a distribution $d \in D$ that $d(i) = 1/2$ and $d(i+1) = 1/2$.

To prove the claim, it is sufficient to show that, if $\mathscr{D}$ is computably PAC learnable wrt the indexes by a probabilistic P-time recursive function (according to the representation $\langle \mathbf{N}, h \rangle$ of course), then $P(x)$ is in RP, therefore $NP = RP$. Let $\chi(\sigma, \tau, n)$ be a probabilistic recursive function, and let $S(x, n), T(y, n)$ be polynomials such that, if σ is chosen at random according to the distribution d_i, if $lth(\sigma) \geqslant S(l(i), n)$ and $lth(\tau) \geqslant T(lth(\sigma), n)$, then with probability $> 1 - 1/n\, \chi(\sigma, \tau, n)$ outputs a j such that $D(d_i, d_j) < 1/n$.

We define a probabilistic recursive function $\Delta(n,\tau,k)$ and a polynomial $U(x,k)$ such that if $lth(\tau)\geqslant U(l(n),k)$, then $\Delta(n,\tau,k)$ decides $P(n)$ in time polynomial in $l(n),k$, with probability $>1-1/k$. $\Delta(n,\tau,k)$ is defined as follows. Suppose $lth(\tau)\geqslant S(l((n,n)^\bullet),k)+T(S(l((n,n)^\bullet),k),n)$. Use the first $S(l((n,n)^\bullet),k)$ bits of τ to simulate $S(l((n,n)^\bullet),k)$ many examples chosen according to a distribution $d_{(n,i)^\bullet}$ s.t. $i\leqslant n$ and $R(n,i)$ (assuming that such an i exists). In other words, let, for $j\leqslant S(l((n,n)^\bullet),k)$, $\sigma_j=n$ if $\tau_j=0$ and $\sigma_j=n+1$ if $\tau_j=1$. Note that if $j,j'\leqslant n$, if $R(n,j)$ and $R(n,j')$ are both true, then $d_{(n,j)^\bullet}=d_{(n,j')^\bullet}$. Then, use the remaining bits to get a binary sequence ρ (obtained by deleting the first $S(l((n,n)^\bullet),k)$ bits of τ) of length $\geqslant T(lth(\sigma),n)$. Finally, compute $\chi(\sigma,\rho,k)$. If $\pi_2(\chi(\sigma,\rho,k))\leqslant n$, check whether $R(n,\pi_2(\chi(\sigma,\rho,k)))$ holds. If so, then $\exists y\leqslant n\ R(n,y)$, therefore $P(n)$ holds. If there exists no distribution $d_{(n,i)^\bullet}$ such that $i\leqslant n$ and $R(n,i)$, χ never outputs h such that $\pi_2(h)\leqslant n$ and $R(n,\pi_2(h))$. Therefore, if $P(n)$ does not hold, then our algorithm gives the correct answer with probability 1. To the other direction, if there is a $j\leqslant n$ such that $R(n,j)$, then σ simulates $S(l((n,n)^\bullet),k)\geqslant S(l((n,j)^\bullet),k)$ examples chosen according to the distribution $d_{(n,j)^\bullet}$. Therefore, with probability $>1-1/k$, $\chi(\sigma,\rho,k)$ is an index h such that $D(d_{(n,j)^\bullet},d_h)<1/k$. If $k>2$, the last inequality implies $D(d_{(n,j)^\bullet},d_h)=0$. Thus we must have $\pi_2(h)\leqslant n$ and $R(n,\pi_2(h))$, i.e. $R(n,\pi_2(\chi(\sigma,\rho,k)))$.

Our probabilistic algorithm works in time polynomial in $l(n),k$, and with probability $>1-1/k$ gives the correct answer to the *NP* question: "does $P(n)$ hold"? □

6. Learning by tests

The frequentist learner fails to identify classes of distributions that are not polynomially localized. Many of such classes are computably PAC learnable wrt the indexes. A trivial example is the following: let

$$\beta=\sum_{n\in\mathbf{N}}\frac{1}{(n+2)\ln^2(n+2)},\quad\text{and let}\quad d(n)=\frac{1}{\beta(n+2)\ln^2(n+2)}.$$

Clearly, the class $\mathscr{D}\equiv\{d\}$ is PAC learnable wrt the indexes (it is sufficient to guess an index for d on any sequence of examples), but not polynomially localized. The problem with the frequentist learner is that he does not take account of the space of hypotheses, and, if the class of distributions is not polynomially localized, he needs too many examples to reach a good approximation of the target distribution. In this section, we present an alternative learning strategy, that takes account of the space of hypotheses.

We start from an example: suppose we are given two distributions d_1, d_2, both not polynomially localized, but such that the probabilities of choosing an even number by a draw according to the distributions d_1 and d_2 are $\frac{1}{3}$ and $\frac{2}{3}$, respectively. Then, the frequentist learner fails to learn $\{d_1,d_2\}$ by Theorem 4.3, even though there is an obvious learning algorithm: count the number E of examples consisting of even

numbers and the number O of examples consisting of odd numbers: if $E<O$, conjecture d_1. Otherwise, conjecture d_2.

Trying to generalize this example, we introduce the notion of *test*. The intuitive idea is that a test for a distribution d is a finite sequence of sets $C_1,\ldots,C_k$ such that we are able to compute approximations of $d(C_1),\ldots,d(C_k)$. Clearly, once we know a test $C_1,\ldots,C_k$ for d, it is reasonable to conjecture d on a sequence σ of examples only if for $i=1\cdots k$ the frequence of C_i relative to σ is sufficiently close to our approximation of $d(C_i)$. It is also intuitively clear that learning by tests is possible whenever for each distribution $d\in\mathscr{D}$ we have a test $C_1,\ldots,C_k$ for d such that, if $d'\in\mathscr{D}$ is such that for $i=1,\ldots,k$ $d'(C_i)$ is very close to $d(C_i)$, then d' is very close to d. In attempting to formalize these intuitive ideas, we introduce the following definitions.

Definition 6.1. Let d be a distribution. Let $n\in\mathbf{N}-\{0\}$. A n test for d is a system of the form $\tau=\langle\langle C_i : i=1,\ldots,k\rangle, g(i)\rangle$ such that for all $i\leqslant k$, C_i is a set of natural numbers, and $|g(i)-d(C_i)|<1/2n$.

Let $\sigma\in Seq$. We say that σ passes the n test τ iff, for all $i\leqslant k$ we have $|f_\sigma(C_i)-g(i)|<1/n$.

Definition 6.2. A uniform P test for a class $\mathscr{D}=\langle d_i : i\in\mathbf{N}\rangle$ of distributions is a system $\mathscr{T}=\langle\langle\langle C^i_{jn} : i\leqslant p(j,n)\rangle : j,n\in\mathbf{N},\ n>0\rangle, g(j,i,n)\rangle$, where $p(j,n)$ is a polynomial, such that the following conditions hold:

- For all $j,n\in\mathbf{N}$, $n>0$, $\langle\langle C^i_{jn} : i\leqslant p(j,n)\rangle, \lambda i.g(j,i,n)\rangle$ is a n test for d_j (called the (j,n)th test of $\mathscr{T}$);
- g is P-time computable;
- The relation $R(j,i,n,x)\equiv i\leqslant p(j,n)$ and $x\in C^i_{jn}$ is in P.

Note that, if we are given a uniform P test $\mathscr{T}$, for all j, n and for all $\sigma\in Seq$, such that $lth(\sigma)\geqslant j$ and $lth(\sigma)\geqslant n$, we can determine by a P-time computation whether or not σ passes the (j,n)th test of $\mathscr{T}$.

Definition 6.3. Let $\mathscr{D}$ and $\mathscr{T}$ be as in Definition 6.2. We say that $\mathscr{T}$ is discriminating for $\mathscr{D}$ iff there is a $k\in\mathbf{N}$ such that for all $j,n\in\mathbf{N}$, $n>0$, the following holds: for all distribution d_r, if for all $i\leqslant p(j,n)|d_r(C^i_{jn})-g(j,i,n)|<2/n$, then $D(d_j,d_r)<\sqrt[k]{1/n}$.

Theorem 6.4. *Assume that there is a uniform P test $\mathscr{T}$ which is discriminating for the class $\mathscr{D}=\langle d_i : i\in\mathbf{N}\rangle$ of distributions. Then $\mathscr{D}$ is computably weakly PAC learnable wrt the indexes.*

Proof. Let $\mathscr{T}=\langle\langle\langle C^i_{jn} : i\leqslant p(j,n)\rangle : j,n\in\mathbf{N},\ n>0\rangle, g(j,i,n)\rangle$. We define a PAC learning algorithm Φ as follows. Given $n\in\mathbf{N}$ and $\sigma\in Seq$ let $\Phi(\sigma,n)$ be the minimal $i\leqslant lth(\sigma)$ such that σ passes the (i,n^k)th test of $\mathscr{T}$, if such an i exists, and 0 otherwise.

Clearly, Φ is a P-time algorithm. We prove that Φ weakly PAC learns $\mathscr{D}$.

Let σ be a sequence drawn at random according to the distribution $d_i \in \mathscr{D}$. We define the events E_0 and E_1 as follows:

$$E_0 : \exists j < i \left(\sigma \text{ passes the } (j, n^k)\text{th test of } \mathscr{T} \text{ and } D(d_i, d_j) \geqslant \frac{1}{n} \right).$$

$$E_1 : \sigma \text{ does not pass the } (i, n^k)\text{th test of } \mathscr{T}.$$

Clearly, if none of E_0, E_1 occurs, then $D(d_i, d_{\Phi(\sigma,n)}) < 1/n$.

Now we bind the probability of E_0. Let $j < i$ be an index such that $D(d_j, d_i) \geqslant 1/n = \sqrt[k]{1/n^k}$. By our assumptions, there is $h \leqslant p(j, n^k)$ such that $|d_i(C^h_{jn^k}) - g(j,h,n^k)| \geqslant 2/n^k$.

Since $|d_i(C^h_{jn^k}) - g(j,h,n^k)| \leqslant |d_i(C^h_{jn^k}) - f_\sigma(C^h_{jn^k})| + |f_\sigma(C^h_{jn^k}) - g(j,h,n^k)|$, if σ passes the (j,n^k)th test of $\mathscr{T}$, then $|d_i(C^h_{jn^k}) - f_\sigma(C^h_{jn^k})| > 1/n^k$. By Chernoff's bound, for any fixed j,h this occurs with probabilty $\leqslant 2\exp(-lth(\sigma)/n^{2k})$.

So, $Pr(E_0) \leqslant 2ip(i,n^k)\exp(-lth(\sigma)/n^{2k})$.

Next, we bind $Pr(E_1)$. If E_1 occurs, then, for some $h \leqslant p(i,n^k)$, one has $|f_\sigma(C^h_{in^k}) - g(i,h,n^k)| \geqslant 1/n^k$. Now $|f_\sigma(C^h_{in^k}) - g(i,h,n^k)| \leqslant |f_\sigma(C^h_{in^k}) - d_i(C^h_{in^k})| + |d_i(C^h_{in^k}) - g(i,h,n^k)|$. Moreover, by the definition of uniform P-test (cf. Definition 6.2), one has $|d_i(C^h_{in^k}) - g(i,h,n^k)| < 1/2(n^k)$. It follows that $|f_\sigma(C^h_{in^k}) - d_i(C^h_{in^k})| > 1/2n^k$. By Chernoff's bounds, for any fixed h, the probability of this is bounded by $2\exp(-lth(\sigma)/4n^{2k})$. So $Pr(E_1) \leqslant 2p(i,n^k)\exp(-lth(\sigma)/4n^{2k})$. We conclude that

$$Pr(E_0 \cup E_1) \leqslant 2ip(i,n^k)\exp\left(-\frac{lth(\sigma)}{n^{2k}}\right) + 2\,p(i,n^k)\exp\left(-\frac{lth(\sigma)}{4n^{2k}}\right).$$

Let $q(i,n) = 2p(i,n^k)(i+1)$, $r(n) = 4n^{2k}$. One has

$$Pr(E_0 \cup E_1) \leqslant q(i,n)\exp\left(-\frac{lth(\sigma)}{r(n)}\right).$$

So, if $lth(\sigma) \geqslant r(n)(\ln(q(i,n)) + \ln(2n))$, one has

$$Pr(E_0 \cup E_1) \leqslant \frac{1}{2n} < \frac{1}{n}.$$

Since if none of E_0, E_1 occurs, then $D(d_i, d_{\Phi(\sigma,n)})$, Φ weakly PAC learns $\mathscr{D}$. □

We may ask whether from the presence of a discriminating uniform P test for the class $\mathscr{D}$ of distributions we can also infer the computable PAC learnability (and not simply weak learnability) of $\mathscr{D}$. The learning algorithm Φ defined in the proof of Theorem 6.4 is P-time, but requires a number of examples polynomial in i (and not

just in $l(i)$) in order to PAC learn d_i. One is tempted to replace Φ by the algorithm Θ defined below.

Definition 6.5. Let $I_{\sigma,n} = \{j \leqslant 2^{lth(\sigma)} : \sigma \text{ passes the } (j,n^k)\text{th test of } \mathscr{T}\}$. We define

$$\Theta(\sigma,n) = \begin{cases} \min(I_{\sigma,n}) & \text{if } I_{\sigma,n} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

The algorithm Θ might fail to be P time for two reasons: the first one is that, since i is only assumed to be bounded by $2^{lth(\sigma)}$ and not simply by $lth(\sigma)$, checking the frequency of all $C^h_{jn^k}$ (a number polynomial in $2^{lth(\sigma)}$) might not be feasible in time polynomial in $l(\sigma^\sharp)$, even if $n \leqslant lth(\sigma)$. We can overcome this difficulty introducing a variant of the concept of uniform P test.

Definition 6.6. A sharply bounded uniform P tests for $\mathscr{D}$ is a system defined as in Definition 6.2 with the only exception that $p(j,n)$ is replaced by $p(l(j),n)$.

Note that, if $\mathscr{T} = \langle\langle\langle C^h_{jn} : h \leqslant p(l(j),n)\rangle : j,n \in \mathbf{N}, n>0\rangle, g(j,h,n)\rangle$, then for all $\sigma \in Seq$ such that $lth(\sigma) \geqslant n$, and for all $j \leqslant 2^{lth(\sigma)}$, checking whether or not σ passes the (j,n^k)th test of $\mathscr{T}$ is polynomial in $lth(\sigma)$, n. Even if we assume that in Definition 6.5 $\mathscr{T}$ is a sharply bounded uniform P test (and not simply a uniform P test), there is another reason why Θ might fail to be P time computable, namely, the presence of a non-logarithmically bounded μ operator (i.e., one bounded by $2^{lth(\sigma)}$). In this respect, we have the same situation we met in Section 5: if $P = NP$ and $\mathscr{T}$ is a sharply bounded uniform discriminating P test for $\mathscr{D}$, then the algorithm Θ defined in Definition 6.5 is P-time and PAC learns $\mathscr{D}$ wrt the indexes. If $NP = RP$, then we can replace Θ by a probabilistic recursive function that works in P-time and PAC learns $\mathscr{D}$. Such a probabilistic recursive function can be replaced by a P time computable function if, in addition, the class $\mathscr{D}$ to be learned is atomless. Proofs are quite similar to those of Theorems 5.7 and 5.9. However, if $NP \neq RP$, then the class of distributions defined in the proof of Theorem 5.10 is not computably PAC learnable, even though it is atomless, and there is a sharply bounded discriminating uniform P test for it.

Summing up, we can conclude

Theorem 6.7. (a) *If $NP = RP$, then every class $\mathscr{D}$ of distributions for which there is a sharply bounded discriminating uniform P test is computably PAC learnable wrt the indexes by a probabilistic recursive function.*

(b) *If $NP = RP$, then every $\mathscr{D}$ as in (a), which, in addition, is atomless, is computably PAC learnable wrt the indexes.*

(c) *If $NP \neq RP$, there is an atomless class $\mathscr{D}$, admitting a sharply bounded discriminating uniform P test, which is not computably PAC learnable wrt the indexes, not even by a probabilistic recursive function.*

References

[1] D. Angluin, Inductive inference of formal languages from positive data, Inform. Control 45 (2) (1980) 117–135.
[2] D. Angluin, Identifying languages from stochastic examples, Technical Report 614, University of Yale, 1988.
[3] S. Ben David, M. Lindenbaum, Learning distributions by their density levels, Lecture Notes in Computer Science, 1996, pp. 53–68.
[4] P. Billingsey, Probability and Measure, 2nd edition, Wiley, New York, 1986.
[5] E.M. Gold, Language identification in the limit, Inform. Control 10 (1967) 447–474.
[6] M.J. Kearnes, Y. Mansour, D. Ron, R.E. Schapire, L. Sellie, On the learnability of discrete distributions, Proc. 26th ACM STOC 94, 1994, pp. 273–282.
[7] M.J. Kearnes, R.E. Schapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. System Sci. 48, 464–497.
[8] M.J. Kearnes, U.V. Vazirani, An Introduction to Computational Learning Theory, MIT Press, Cambridge, MA, 1994.
[9] F. Montagna, Investigations in measure-one identification of classes of languages, Inform. Comput. 143, 74–107.
[10] B.K. Natarajan, Probably approximate learning over classes of distributions, SIAM J. Comput. 21, 438–449.
[11] P. Odifreddi, Classical Recursion Theory, North-Holland, Amsterdam, 1989.
[12] P. Odifreddi, Classical Recursion Theory, Vol II, book in preparation.
[13] D. Osherson, M. Stob, S. Weinstein, Systems that Learn, An Introduction to Learning Theory for Cognitive and Computer Scientists, MIT Press, Cambridge, MA, 1986.
[14] L. Pitt, Probabilistic inductive inference, J. ACM 36 (2) (1989) 383–433.
[15] L. Pitt, C. Smith, Probability and plurality for aggregations of learning machines, Inform. Comput. 77 (1988) 77–92.
[16] L.G. Valiant, A theory of learnable, Comm. ACM 27 (11) (1984) 1134–1142.