Polynomial Uniform Convergence and Polynomial-Sample Learnability*

Alberto Bertoni†, Paola Campadelli†,
Anna Morpurgo†‡, and Sandra Panizza†

TR-92-077

November 1992

Abstract

In the PAC model, polynomial–sample learnability in the distribution dependent framework has been characterized in terms of minimun cardinality of ϵ-covers. In this paper we propose another approach to the problem by investigating the relationship between polynomial–sample learnability and uniform convergence, in analogy to what was done for the distribution free setting. First of all, we introduce the notion of polynomial uniform convergence, giving a characterization for it in terms of an entropic measure, then we study its relationship with polynomial–sample learnability. We show that, contrarily to what happens in the distribution independent setting, polynomial uniform convergence is a sufficient but not necessary condition for polynomial–sample learnability.

*This research was partly supported by CNR, grant 92.01568.PF.69, project *Sistemi Informatici e Calcolo Parallelo*. An extended abstract of this paper appeared in *Proc. 5th Annual ACM Workshop on Computational Learning Theory* (1992).

†Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy. Email addresses: bertoni@hermes.mc.dsi.unimi.it, campadelli@hermes.mc.dsi.unimi.it, morpurgo@imiucca.csi.unimi.it, panizza@hermes.mc.dsi.unimi.it.

‡Part of this research was performed while the author was visitor at the International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94704.

1 Introduction

The *probably approximately correct (PAC) learning model* proposed by Valiant [Va84] provides a complexity theoretical basis to study the problem of learning from examples produced by an arbitrary distribution.

The model can be informally described as follows. Given a domain X, a *concept class* C is a set of subsets of X. In learning a class of concepts C from examples, a *target concept* is selected from C and we are given a finite sequence of elements of X, each labelled "1" if it is in the target concept (a positive example) and "0" otherwise. This set is called a *labelled sample* of the target concept. Let P be a fixed, but arbitrary and unknown, probability distribution on X; the examples are created by drawing points independently at random according to P. A *learning algorithm* for C is an algorithm that, for any distribution P, given a large enough sample randomly drawn according to P of any target concept in C, returns a set in X (a *hypothesis*) that is with high probability a good approximation to the target concept. A class C for which there exists a learning algorithm is called *learnable*. For the purposes of computation we must assume some representation for the hypotheses produced by a learning algorithm. In this paper we will consider Boolean concepts. We can assume that the hypotheses produced by the learning algorithm are represented by the Boolean circuit realizing the corresponding function.

Central concepts in this approach to the study of learning are the *sample* and *time complexity* of the learning algorithm, that is the smallest sample size and the minimum number of steps required by the algorithm to produce (the representation of) a hypothesis for any target concept in C and for all distributions on the domain, with given accuracy and confidence.

Valiant's original PAC learning model requires that the algorithm work for any probability distribution on the domain. An interesting restriction of the basic model of Valiant is that of fixing the probability distribution. Benedek and Itai [BeIt88] study the problem with a geometrical approach based on the fact that the probability distribution induces a metric on the error, that is on the probability that the hypothesis disagrees with the target concept, and obtain bounds on the sample complexity in this model in terms of ϵ-covers. They propose a learning algorithm, the best–agreement learning algorithm, which returns as hypothesis one of the elements of the ϵ-cover that has the smallest number of inconsistencies with the sample. They give a nice characterization of distribution dependent learnability in terms of ϵ-covers.

Vapnik and Chervonenkis [VaCh71] studied the convergence of empirical probability estimates, they considered the following problem, that arises for example in pattern recognition: given a class of events, do their relative frequencies in a sequence of independent trials converge to the probabilities uniformly over the entire class of events? More precisely, under what conditions does the probability that the maximum difference (over the class) between the relative frequency and the probability exceeds a given arbitrarily small positive constant tend to zero as the number of trials is increased indefinitely? For the distribution–free setting they introduce a combinatorial measure, called the *Vapnik–Chervonenkis dimension*, which in some sense is a measure of the "dichotomization" power of a class, and give sufficient conditions for uniform convergence. For the distribution–dependent framework they give sufficient and necessary conditions for such convergence and their results are based on

a parameter they call *entropy*, which is a measure of the average information provided by a sample of fixed size.

Blumer et al. [BlEhHaWa89] apply the results obtained by Vapnik and Chervonenkis in the distribution–free setting to the field of computational learning theory and show that the Vapnik–Chervonenkis dimension is a useful notion to characterize the distribution–free learnability of a class of concepts. Moreover they show that the following property holds: in the distribution free framework if a class is learnable then it is learnable by any algorithm which always outputs a hypothesis consistent with the sample.

What about learning in the distribution–dependent model? Can Vapnik and Chervonenkis' results be fruitfully applied also in this context? And what is the exact relationship between uniform convergence and learnability in the two frameworks, the distribution–free and the distribution–dependent one?

In this paper we explore the problem of distribution–dependent learnability studying its relationship with uniform convergence, instead of following the approach proposed in [BeIt88] (Natarajan [Na88] [Na92] also studied the problem with an approach similar to ours); in particular we investigate polynomial–sample learnability, i.e. learnability with a polynomial bound on the sample size, in terms of a parameter tightly related to the one introduced in [VaCh71] for studying the problem of uniform convergence in the distribution–dependent framework. We will restrict our discussion to the Boolean domain, although our results also hold for infinite domains, such as E^n, provided certain measurability constraints on the concept class are satisfied. To investigate the asymptotic complexity of learning from examples we consider families of concept classes. A family of concept classes $\{F_n\}_{n\geq 1}$ is a set of concept classes such that for each $n \geq 1$, F_n is a concept class on the n Boolean variables $x_1, \ldots, x_n$. We introduce the concept of *polynomial uniform convergence* of relative frequencies to probabilities: consider, for any n, the domain $X_n = \{0,1\}^n$, a probability distribution on X_n, and a set of events $F_n \subseteq 2^{X_n}$; we say that the class F_n is polynomially uniformly convergent if the probability that the maximum difference (over F_n) between the relative frequency and the probability of an event exceed a given arbitrarily small positive constant ϵ is at most δ $(0 < \delta < 1)$ when the sample on which we estimate the relative frequencies has size polynomial in $n, 1/\epsilon, 1/\delta$. We present a necessary and sufficient condition for polynomial uniform convergence in terms of an entropic measure, we then prove that this condition is sufficient but not necessary for polynomial–sample learnability, contrarily to what happens for the distribution–independent framework, where polynomial uniform convergence is a necessary and sufficient condition for polynomial–sample learnability.

In section 2 we introduce the concept of polynomial uniform convergence in the distribution–free framework. In section 3 we discuss the relationship between polynomial uniform convergence and polynomial–sample learnability in the distribution–free framework, and give a general characterization of both in terms of the Vapnik–Chervonenkis dimension. In section 4 we introduce an entropic measure related to the one introduced in [VaCh71], consider the concept of polynomial uniform convergence in the distribution–dependent framework and characterize it in terms of this entropic measure. The relationship between polynomial uniform convergence and polynomial–sample learnability in the distribution dependent framework is then discussed in section 5.

Notation. The cardinality of a set S will be denoted by $\#S$, the simmetric difference

between two set S and T, $S-T \cup T-S$, will be denoted by $S \triangle T$, the symbol $\wedge$ will denote the logic operator AND, the base 2 logarithm will be denoted by log.

2 Preliminary Definitions and Results

In this section we recall the PAC learning model, present the results in [VaCh71] on uniform convergence, proving a stronger version of one of them, and introduce the notion of polynomial uniform convergence.

Let us first give the basic notation. Let X be a set of elementary events on which a probability measure P is defined and let F be a collection of boolean functions on X, i.e. functions $f : X \to \{0,1\}$. For $f \in F$ the set $f^{-1}(1)$ is said event, and $\mathcal{P}_f$ denotes its probability. A *t-sample* (or sample of size t) on X is a sequence $\underline{x} = (x_1, \ldots, x_t)$, where $x_k \in X$ $(1 \leq k \leq t)$. Let $X^{(t)}$ denote the space of t-samples and $P^{(t)}$ the probability distribution induced by P on $X^{(t)}$, such that $P^{(t)}(x_1, \ldots, x_t) = P(x_1)P(x_2)\cdots P(x_t)$.

Given a t-sample $\underline{x}$ and a function $f \in F$, let $\nu_f(\underline{x})$ be the relative frequency of f in the t-sample $\underline{x}$, i.e.

$$\nu_f(\underline{x}) = \frac{\sum_{i=1}^{t} f(x_i)}{t}.$$

Consider now the random variable $\pi_F^{(t)} : X^{(t)} \to [0,1]$, defined over $\langle X^{(t)}, P^{(t)} \rangle$, where

$$\pi_F^{(t)}(x_1, \ldots, x_t) = \sup_{f \in F} \mid \nu_f(x_1, \ldots, x_t) - \mathcal{P}_f \mid .$$

The relative frequencies of the events are said to *converge uniformly* over F to the probabilities if, for every $\epsilon > 0$, $\lim_{t\to\infty} P^{(t)}\{\underline{x} \mid \pi_F^{(t)}(\underline{x}) > \epsilon\} = 0$.

We now introduce the PAC learning model proposed by Valiant [Va84] and its distribution–dependent restriction.

For our purposes it is convenient to consider parametrized domains and classes of concepts. Let $X_n = \{0,1\}^n$ and let F_n be a class of boolean functions on X_n; a labelled t-sample S_f for $f \in F_n$ is a sequence $(\langle x_1, f(x_1)\rangle, \ldots, \langle x_t, f(x_t)\rangle)$, where $(x_1, \ldots, x_t)$ is a t-sample on X_n. Fixed a probability distribution P_n on X_n, we say that $f_1, f_2 \in F_n$ are *ϵ-close* with respect to P_n iff $P_n\{x | f_1(x) \neq f_2(x)\} \leq \epsilon$.

We consider the learnability of families of concept classes. An algorithm A is a *probably approximately correct (PAC) learning algorithm* with sample size $t(n, 1/\epsilon, 1/\delta)$ for the family $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ if, given in input ϵ, δ and n, the following holds: for any $n \geq 1$, for any function $f \in F_n$, for any probability distribution P_n over X_n, and for any $0 < \epsilon, \delta < 1$, if the algorithm A makes at most $t(n, 1/\epsilon, 1/\delta)$ calls to an oracle which at every call provides an example of f drawn independently at random according to P_n, then with probability at least $1 - \delta$ it outputs as hypothesis a representation of a function g that is ϵ-close to f. If furthermore the algorithm always outputs a hypothesis that is consistent with the target on the examples seen, then it is called a *consistent learning algorithm.*

Definition 2.1 *A family* $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ *is* polynomial–sample learnable *if there exists a learning algorithm A for it with sample size $t(n, 1/\epsilon, 1/\delta)$ bounded by some polynomial in $1/\epsilon, 1/\delta, n$.*

By relaxing the condition that a learning algorithm work regardless of the distribution on the examples, the model is considered under fixed distribution. Formally, we consider a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, where P_n is a probability distribution on X_n; in this case the definition of polynomial–sample learnability is the same as before, except that the performance criteria for learnability must be met only under the fixed distribution.

We will now recall the notion of Vapnik–Chervonenkis dimension and prove that the sufficient condition for uniform convergence given in [VaCh71] for the distribution–free case is also a necessary one. We will then present Vapnik and Chervonenkis' result for the distribution–dependent case.

In order to study the problem of uniform convergence of the relative frequencies to the probabilities the notion of index $\Delta_F(\underline{x})$ of the class F with respect to a t-sample $\underline{x}$ has been introduced [VaCh71]. Fixed the t-sample $\underline{x} = (x_1, \ldots, x_t)$,

$$\Delta_F(\underline{x}) = \#\{f^{-1}(1) \cap \{x_1, \ldots, x_t\} | f \in F\}.$$

Obviously $\Delta_F(x_1, \ldots, x_t) \leq 2^t$; a set $\{x_1, \ldots, x_t\}$ is said *shattered* by F iff $\Delta_F(x_1, \ldots, x_t) = 2^t$; the maximum t such that there is a set $\{x_1, \ldots, x_t\}$ shattered by F is said the *Vapnik–Chervonenkis dimension (VCdim)* of F. This combinatorial parameter is very important in the distribution independent context. In particular the following result holds.

Theorem 2.1 *For all probability distributions on X, the relative frequencies of the events converge (in probability) to their probabilities uniformly over F iff VCdim(F) $< \infty$.*

Proof. The if part is proved in [VaCh71]. Assume now that the relative frequencies of the events converge (in probability) to their probabilities uniformly over F, that is for every $\epsilon > 0$, $\lim_{t\to\infty} P^{(t)}\{\underline{x} \mid \pi_F^{(t)}(\underline{x}) > \epsilon\} = 0$. Consider the class $f\triangle F = \{f\triangle h \mid h \in F\}$ of the simmetric differences between f and the functions in F, for this class also holds

$$\lim_{t\to\infty} P^{(t)}\{\underline{x} \mid \pi_{f\triangle F}^{(t)}(\underline{x}) > \epsilon\} = 0. \tag{1}$$

Let $F_{f,\underline{x}}$ denote the subclass of F of the functions f' that are consistent with f on the sample $\underline{x}$. Trivially

$$P^{(t)}\{\underline{x} \mid \pi_{f\triangle F}^{(t)}(\underline{x}) > \epsilon\} \geq P^{(t)}\{\underline{x} \mid \pi_{f\triangle F_{f,\underline{x}}}^{(t)}(\underline{x}) > \epsilon\}. \tag{2}$$

Besides $\nu_{f\triangle f'}(\underline{x}) = 0$ for all $f' \in F_{f,\underline{x}}$, which yields

$$P^{(t)}\{\underline{x} \mid \pi_{f\triangle F_{f,\underline{x}}}^{(t)}(\underline{x}) > \epsilon\} = P^{(t)}\{\underline{x} \mid \mathcal{P}_{f\triangle f'} > \epsilon, f' \in F_{f,\underline{x}}\}, \tag{3}$$

where the last expression is the probability that a consistent hypothesis has error greater than ϵ. Thus, if for a class F the relative frequencies of the events in F converge to their probabilities uniformly over F, then by (1), (2), and (3) the class is learnable by consistent hypotheses (solidly learnable [BeBeMa89]). As shown in [BlEhHaWa89], a class is solidly learnable iff it has finite VC dimension, which completes the proof. □

In the distribution–dependent context, where the probability measure P is fixed and known, the problem of uniform convergence has been characterized in terms of the expectation $E[\log_2 \Delta_F(\underline{x})]$, called entropy $H_F(t)$ of the class F in samples of size t. In fact the following result [VaCh71] holds.

Theorem 2.2 *A necessary and sufficient condition for the relative frequencies of the events in F to converge uniformly over F (in probability) to their corresponding probabilities is that*

$$\lim_{t\to\infty} \frac{H_F(t)}{t} = 0.$$

Vapnik and Chervonenkis [VaCh71] studied the problem of uniform convergence of the frequencies to the probabilities in the limit. In order to study the relationship between the concepts of uniform convergence and polynomial–sample learnability we need to analize the rate of convergence of the frequencies to the probabilities in a class, we thus introduce the concept of *polynomial uniform convergence*.

Definition 2.2 *A family $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ is said to be* polynomially uniformly convergent *iff there exists a polynomial $p(n, 1/\epsilon, 1/\delta)$ such that, for all n and all probability distributions P_n on X_n,*

$$\forall \epsilon, \delta > 0\, (t \geq p(n, 1/\epsilon, 1/\delta) \Rightarrow P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta).$$

The problem of learnability in the distribution–dependent framework has been studied by Benedek and Itai [BeIt88] with a different approach using the notion of *coverability*. Let F_n be a class of functions on X_n and P_n a probability distribution on X_n. An *ϵ-cover* of F_n w.r.t. P_n is a set $F_{n,\epsilon} \subseteq 2^{X_n}$ such that for every $f \in F_n$ there is an $\tilde{f} \in F_{n,\epsilon}$ ϵ-close to f. This notion plays a role analogous to the Vapnik–Chervonenkis dimension, in the sense that it characterizes learnability in the distribution–dependent framework. In fact from the results in [BeIt88] readily follows that a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is polynomial–sample learnable iff there is a polynomial $p(n, 1/\epsilon)$ such that

$$N_{P_n}(F_n, \epsilon) \leq 2^{p(n,1/\epsilon)},$$

where $N_{P_n}(F_n, \epsilon)$ is the cardinality of the smallest ϵ-cover of F_n w.r.t. P_n.

3 Polynomial uniform convergence and polynomial–sample learnability: distribution–independent case

In this section we study the relationship between polynomial uniform convergence and polynomial–sample learnability in the distribution–independent framework. An important characterization of polynomial–sample learnability has been given in terms of the Vapnik–Chervonenkis dimension; in fact it has been shown [BlEhHaWa89] that a family $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ is polynomial–sample learnable iff the Vapnik–Chervonenkis dimension of F_n is bounded by a polynomial in n.

We strengthen their result and show that both the polynomial–sample learnability and the polynomial uniform convergence of a family $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ can be characterized in terms of the Vapnik–Chervonenkis dimension. In fact the following theorem holds.

Theorem 3.1 *Given a family $\{\langle X_n, F_n\rangle\}_{n\geq 1}$, the following conditions are equivalent:*

C1. *The relative frequencies of events in F_n converge uniformly polynomially to their probabilities.*

C2. $\{\langle X_n, F_n\rangle\}_{n\geq 1}$ *is polynomial–sample learnable.*

C3. $d_n = VCdim(F_n)$ *is bounded by a polynomial in* n.

Proof.

- C2 $\Leftrightarrow$ C3 is proved in [BlEhHaWa89].
- C3 $\Rightarrow$ C1 is an immediate consequence of the results in [VaCh71]. In fact there it is proved that if $t \geq \frac{2}{\epsilon^2}$ then

$$P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} \leq 4((2t)^{d_n+1}+1)e^{-\frac{\epsilon^2 t}{8}}$$

  and that, if $t = \frac{16}{\epsilon^2}((d_n+1)\log\frac{16(d_n+1)}{\epsilon^2} - \log\frac{\delta}{4})$, then $P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta$. By hypothesis d_n is bounded by a polynomial in n, hence the thesis follows.

- C1 $\Rightarrow$ C3. Note that, since d_n is finite, there exists a set A of d_n points $\{a_1, \ldots, a_{d_n}\} \subseteq X_n$ that is shattered by F_n. Let the probability distribution Q_n on X_n be uniform on these points and 0 elsewhere. Given a t-sample $(x_1, \ldots, x_t)$ such that $Q_n^{(t)}(x_1, \ldots, x_t) \neq 0$ and $t \leq \frac{1}{2}d_n$ we have, denoting $\{x_1, \ldots, x_t\}$ with s,

$$\pi_{F_n}^{(t)}(x_1, \ldots, x_t) = \sup_{f\in F_n} |\nu_f(x_1, \ldots, x_t) - \mathcal{Q}_f| \geq |\nu_s(x_1, \ldots, x_t) - \mathcal{Q}_s| \geq 1 - \frac{t}{d_n} > \frac{1}{3};$$

  therefore

$$t \leq \frac{1}{2}d_n \Rightarrow Q_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > 1/3\} = 1 > 1/2. \tag{4}$$

  By hypothesis there exists a polynomial $p(n, 1/\epsilon, 1/\delta)$ such that, for all ϵ, $\delta > 0$, for all $n \geq 1$, and for every probability distribution P_n on X_n,

$$t \geq p(n, 1/\epsilon, 1/\delta) \Rightarrow P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta.$$

  Now letting $\epsilon = 1/3$, $\delta = 1/2$ and considering the particular distribution Q_n, assertion (4) implies that $d_n < 2p(n, 3, 2)$. □

4 Polynomial uniform convergence: distribution–dependent case

The distribution–free framework is in some sense the most desirable one because the positive results proven in it are the strongest. In real life situations one often has information about the probability distribution on the domain, moreover many interesting classes are not learnable in the distribution–free framework, but some of them are learnable under some specific distribution. A model of learnability naturally arises, where the probability distribution on the domain is fixed and known.

In the distribution–dependent context the definition of polynomial uniform convergence is the same as that of Definition 2.2 except that the convergence criteria must be met only for the fixed distribution.

We will now study the problem of polynomial uniform convergence in the distribution–dependent framework and will see that the conditions are not the same as for Theorem 3.1, in fact we need stronger conditions for C1. More precisely we consider the problem of characterizing the families $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, where P_n is a fixed and known probability distribution on X_n, such that the relative frequencies of the events in F_n converge polynomially to their probabilities and we will give such a characterization in terms of an entropical parameter.

Let us introduce the random variable $\mathcal{C}_n^{(t)} : X_n^{(t)} \to N$, defined as

$$\mathcal{C}_n^{(t)}(x_1, \ldots, x_t) = \max\{\#A \mid A \subseteq \{x_1, \ldots, x_t\} \wedge A \text{ is shattered by } F_n\}.$$

In this notation it is understood that $\mathcal{C}_n^{(t)}$ refers to F_n. The random variable $\mathcal{C}_n^{(t)}$ and the index function Δ_{F_n} are related to one another; in fact, the following result holds [Pa91].

Lemma 4.1 $\mathcal{C}_n^{(t)}(\underline{x}) \leq \log \Delta_{F_n}(\underline{x}) \leq \log(t^{\mathcal{C}_n^{(t)}(\underline{x})} + 1)$.

Proof. The left inequality holds trivially. For the right inequality the case $\mathcal{C}_n^{(t)} = t$ is also trivial. If $\mathcal{C}_n^{(t)} < t$, then no subset of $\{x_1, \ldots, x_t\}$ of cardinality $\mathcal{C}_n^{(t)} + 1$ is shattered by F_n. As shown in [Sa72],

$$\Delta_{F_n}(\underline{x}) \leq \Phi(\mathcal{C}_n^t(\underline{x}), t),$$

where

$$\Phi(n,t) = \begin{cases} \sum_{k=0}^{n} \begin{pmatrix} t \\ k \end{pmatrix} & \text{if } t > n \\ 2^t & \text{if } t \leq n \end{cases}$$

From the known bound

$$\Phi(n,t) \leq t^n + 1$$

we get

$$\Delta_{F_n}(\underline{x}) \leq t^{\mathcal{C}_n^t(\underline{x})} + 1$$

which proves the right inequality of the lemma. □

For $t \geq 2$ the relation

$$\mathcal{C}_n^{(t)}(\underline{x}) \leq \log \Delta_{F_n}(\underline{x}) \leq \mathcal{C}_n^{(t)}(\underline{x}) \cdot \log t \tag{5}$$

holds, so from now on we will use this simpler relation without further specifing that it holds for $t \geq 2$.

Let $M(n,t) = \mathrm{E}(\mathcal{C}_n^{(t)}/t)$ be the expectation of the random variable $\mathcal{C}_n^{(t)}/t$. From Lemma 4.1 (relation (5)) readily follows that

$$M(n,t) \leq \frac{H_{F_n}(t)}{t} \leq M(n,t) \cdot \log t;$$

therefore $M(n,t)$ is very close to $H_{F_n}(t)/t$, which can be interpreted as average information per example for samples of size t.

$M(n,t)$ is a useful measure to verify whether $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ satisfies the property of polynomial uniform convergence, as we will show in Theorem 4.3. In order to prove Theorem 4.3 we need the following results.

The following lemma is obtained by minor modifications from [VaCh71, Lemma 2, Theorem 4, and Lemma 4]. (A similar result was independently proved in [Na92].)

Lemma 4.2 *Given* $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, *if* $\lim_{t\to\infty} H_{F_n}(t)/t = 0$ *then*

$$\forall\epsilon\ \forall\delta\ \forall n(t \geq \frac{128t_0}{\epsilon^2\delta} \Rightarrow P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta),$$

where t_0 *is such that* $H_{F_n}(t_0)/t_0 \leq \epsilon^2/64$.

Proof. Let $\underline{x} \in X_n^{(t)}$ and $\underline{y} \in X_n^{(2t)}$ and let $\underline{x}'$ and $\underline{x}''$ denote the first and the second half, respectively, of $\underline{y}$. In [VaCh71, Lemma 2] it is proved that, if $t > 2/\epsilon^2$, then

$$P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} \leq 2P_n^{(2t)}\{\underline{y} \mid \sup_{f\in F_n} |\nu_f(\underline{x}') - \nu_f(\underline{x}'')| \geq \frac{\epsilon}{2}\}. \tag{6}$$

Let us denote the event $\{\sup_{f\in F_n} |\nu_f(\underline{x}') - \nu_f(\underline{x}'')| \geq \frac{\epsilon}{2}\}$ by C and estimate $P_n^{(2t)}\{C\}$. It is convenient to introduce the random variable

$$\xi_n^{(t)}(\underline{x}) = \frac{\log \Delta_{F_n}(x_1, \ldots, x_t)}{t}$$

and to denote $P_n^{(t)}\{\underline{x} \mid \xi_n^{(t)}(\underline{x}) > \epsilon\}$ by $P^+(t,\epsilon)$. As proved in [VaCh71],

$$P_n^{(2t)}\{C\} \leq 2(2/e)^{\frac{\epsilon^2 t}{8}} + P^+(2t, \frac{\epsilon^2}{16}).$$

Let $\alpha = \epsilon^2/16$. Let t_0 be such that $H_{F_n}(t_0)/t_0 \leq \alpha/4$ (such a t_0 must exist because by hypothesis $\lim_{t\to\infty} H_{F_n}(t)/t = 0$). Following [VaCh71], let us first estimate $P^+(t, \alpha/2)$ with $t = mt_0$ (m integer) and obtain

$$P^+\left(mt_0, \frac{\alpha}{2}\right) \leq \frac{16\ var(\xi_n^{(t_0)})}{m\alpha^2}. \tag{7}$$

From (7) let us now obtain a bound on $P^+(2t,\alpha)$, where $\alpha = \epsilon^2/16$ and t is arbitrary. For arbitrary $2t > t_0$ and m such that $mt_0 < 2t < (m+1)t_0$, where t_0 is such that $H_{F_n}(t_0)/t_0 < \alpha/4$,

$$\frac{1}{mt_0}\log\Delta(x_1, \ldots, x_{(m+1)t_0}) > \frac{1}{2t}\log\Delta(x_1, \ldots, x_{2t}).$$

Thus

$$P\left\{\frac{m+1}{m}\xi_n^{((m+1)t_0)} > \alpha\right\} > P^+(2t,\alpha). \tag{8}$$

For m sufficiently large,

$$P\left\{\frac{m+1}{m}\xi_n^{((m+1)t_0)} > \alpha\right\} \leq P\left\{\xi_n^{((m+1)t_0)} > \frac{\alpha}{2}\right\} = P^+\left((m+1)t_0, \frac{\alpha}{2}\right). \tag{9}$$

From (8), (9), and (7) readily follows that

$$P^+(2t,\alpha) < P^+\left((m+1)t_0, \frac{\alpha}{2}\right) \leq \frac{16\, var(\xi_n^{(t_0)})}{(m+1)\alpha^2} \leq \frac{16\, t_0\, var(\xi_n^{(t_0)})}{2t\alpha^2}$$

Letting $\alpha = \epsilon^2/16$ we obtain

$$P_n^{(2t)}\{C\} < 2(2/e)^{\frac{\epsilon^2 t}{8}} + \frac{16^3\, t_0\, var(\xi_n^{(t_0)})}{2t\epsilon^4} \tag{10}$$

Since $0 \leq var(\xi_n^{(t_0)}) \leq H_{F_n}(t_0)/t_0$, we have $var(\xi_n^{(t_0)}) \leq \epsilon^2/64$. Substituting in (10) yields

$$P_n^{(2t)}\{C\} < 2(2/e)^{\frac{\epsilon^2 t}{8}} + \frac{32 t_0}{t\epsilon^2}.$$

By substituting $P_n^{(2t)}\{C\}$ with the above bound in (6) we obtain

$$P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} \leq 4(2/e)^{\frac{\epsilon^2 t}{8}} + \frac{64 t_0}{t\epsilon^2}. \tag{11}$$

We want to estimate the sample size $t(n, 1/\epsilon, 1/\delta)$ such that $P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta$. It suffices to find a t such that both

$$4(2/e)^{\frac{\epsilon^2 t}{8}} < \frac{\delta}{2}$$

and

$$\frac{64 t_0}{t\epsilon^2} < \frac{\delta}{2}$$

hold. We obtain

$$t > \max\left\{\frac{8}{\epsilon^2}\frac{\log(8/\delta)}{\log(e/2)}, \frac{16 t_0}{\epsilon^2}\frac{8}{\delta}, \frac{2}{\epsilon^2}\right\}$$

which yields

$$t > \frac{128 t_0}{\epsilon^2 \delta}.$$

□

Lemma 4.2 states that a family $\{\langle X_n\, ,\, P_n\, ,\, \mathcal{F}_n\rangle\}_{n\geq 1}$ converges uniformly polynomially if the sample size t_0 is such that $H_{F_n}(t_0)/t_0 \leq \epsilon^2/64$ and it is polynomial in $1/\epsilon$ e n. Formally:

Theorem 4.1 *Given $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, if there exists a polynomial $\psi(n, \frac{1}{\epsilon})$ such that*

$$\forall\epsilon\, \forall n(t \geq \psi(n, \frac{1}{\epsilon}) \Rightarrow \frac{H_{F_n}(t)}{t} \leq \epsilon),$$

then the relative frequencies of the events in F_n converge polynomially to their probabilities.

Proof (outline). It is sufficient to observe that if we choose $t_0 = \psi(n, 64/\epsilon^2)$, by hypothesis it holds that $H_{F_n}(t_0)/t_0 \leq \epsilon^2/64$; therefore, from Lemma 4.2, if

$$t \geq \frac{128 t_0}{\epsilon^2\delta} = \frac{128}{\epsilon^2\delta}\,\psi(n, \frac{64}{\epsilon^2}),$$

then $P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} < \delta$. □

The following lemma, which relates the problem of polynomial uniform convergence of a family of events to the parameter $P_a(n, \epsilon, t)$, will only be stated since it is proved by minor modifications of Theorem 4 in [VaCH71]. For the sake of simplicity it is convenient to introduce the following notations:

$$a_n^{(t)} = \mathcal{C}_n^{(t)}/t \qquad P_a(n, \epsilon, t) = P_n^{(t)}\{\underline{x} \mid a_n^{(t)}(\underline{x}) \leq \epsilon\}.$$

Lemma 4.3 *If* $t \geq 16/\epsilon^2$ *then* $P_n^{(t)}\{\underline{x} \mid \pi_{F_n}^{(t)}(\underline{x}) > \epsilon\} \geq \frac{1}{4}(1 - P_a(n, 8\epsilon, 2t))$.

A relevant property of $P_a(n, \epsilon, t)$ is given by the following lemma.

Lemma 4.4 $\forall\, \alpha \geq 1 \;\; P_a(n, \frac{\epsilon}{\alpha}, \alpha t) \leq P_a^{\alpha}(n, \epsilon, t)$.

Proof. Let $(\underline{x}_1, \ldots, \underline{x}_\alpha)$ be an αt_sample obtained by the concatenation of α elements $\underline{x}_1, \ldots, \underline{x}_\alpha \in X^{(t)}$. It is easy to verify that $\mathcal{C}_n^{(\alpha t)}(\underline{x}_1, \ldots, \underline{x}_\alpha) \geq \max_{i=1,\ldots,\alpha} \mathcal{C}_n^{(t)}(\underline{x}_i)$. Therefore

$$P_n^{(\alpha t)}\{\mathcal{C}_n^{(\alpha t)}(\underline{x}_1, \ldots, \underline{x}_\alpha) \leq k\} \leq P_n^{(\alpha t)}\{\mathcal{C}_n^{(t)}(\underline{x}_1) \leq k \wedge \cdots \wedge \mathcal{C}_n^{(t)}(\underline{x}_\alpha) \leq k\}.$$

By the independency of the events $\mathcal{C}_n^{(t)}(\underline{x}_i) \leq k$ we obtain

$$P_n^{(\alpha t)}\{\mathcal{C}_n^{(\alpha t)}(\underline{x}_1, \ldots, \underline{x}_\alpha) \leq k\} \leq \prod_{i=1}^{\alpha} P_n^{(t)}\{\mathcal{C}_n^{(t)}(\underline{x}_i) \leq k\}.$$

Recalling that $a_n^{(t)} = \mathcal{C}_n^{(t)}/t$ and substituting $k = \epsilon t$, the thesis follows. 3pt □

A relation between $P_a(n, \epsilon, t)$ and the parameter $M(n, t)$, which we have introduced to characterize the polynomial uniform convergence of $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, is shown in the following lemma.

Lemma 4.5 *For every* ϵ $(0 < \epsilon < 1/4)$, *if* $M(n,t) > 2\sqrt{\epsilon}$ *then* $P_a(n, \epsilon, t) < 1/2$.

Proof. For the sake of simplicity, let $m = M(n, l)$. If $m > \delta > 0$, we have

$$\delta < m = \int_0^1 x\, dP_a = \int_0^{\delta/2} x\, dP_a + \int_{\delta/2}^1 x\, dP_a \leq \frac{\delta}{2} P_a(n, \frac{\delta}{2}, l) + 1 - P_a(n, \frac{\delta}{2}, l).$$

Since $0 < \delta < 1$, we obtain $P_a(n, \frac{\delta}{2}, l) < \dfrac{1-\delta}{1-\delta/2} \leq 1 - \dfrac{\delta}{2}$. By applying Lemma 4.4 it is proved that, for every $\alpha \geq 1$,

$$P_a(n, \frac{\delta}{2\alpha}, \alpha l) \leq \left(1 - \frac{\delta}{2}\right)^{\alpha}.$$

For $\alpha = \dfrac{2}{\delta}$ we obtain $P_a(n, \dfrac{\delta^2}{4}, \dfrac{2l}{\delta}) < e^{-1} < \dfrac{1}{2}$. For $\epsilon = \delta^2/4$ and $t = 2l/\delta$ the previous result implies that, if $M(n, t\sqrt{\epsilon}) > 2\sqrt{\epsilon}$, then $P_a(n, \epsilon, t) < 1/2$.

It is easy to verify that $\mathcal{C}_n^{(\alpha t)}(\underline{x}_1, \ldots, \underline{x}_\alpha) \leq \sum_{i=1}^{\alpha} \mathcal{C}_n^{(t)}(\underline{x}_i)$ for every $\alpha \geq 1$. This implies $M(n, \alpha t) \leq M(n, t)$ for $\alpha \geq 1$, hence $M(n, t\sqrt{\epsilon}) \geq M(n, t)$, from which the thesis follows. □

Theorem 4.2 *If for the family* $\{\langle X_n\,,\,P_n\,,\,F_n\rangle\}_{n\geq 1}$ *the relative frequencies of events in* F_n *converge polynomially to their probabilities then there exists a polynomial* $\psi(n,1/\epsilon)$ *such that*

$$\forall\,\epsilon\;\forall n\,(t\geq\psi(n,1/\epsilon)\Rightarrow M(n,t)\leq\epsilon).$$

Proof. By contradiction. Let us suppose that $\{\langle X_n,P_n,F_n\rangle\}_{n\geq 1}$ polynomially converges and that for all polynomial functions $\psi(n,\frac{1}{\epsilon})$ there exist ϵ, n, t such that $t\geq\psi(n,\frac{1}{\epsilon})$ and $M(n,t)>\epsilon$.

Since $M(n,t)$ is a monotone non increasing function with respect to t, it follows that for every ψ there exist ϵ, n such that $M(n,\psi(n,\frac{1}{\epsilon}))>\epsilon$. Considering the one–to–one corrispondence T between polynomial functions defined by $T\psi(n,\frac{1}{\epsilon})=\varphi(n,\frac{4}{\epsilon^2})$, we can conclude that for all φ there exist ϵ, n such that $M(n,\varphi(n,\frac{1}{\epsilon}))>2\sqrt{\epsilon}$. By applying Lemma 4.5 we obtain

$$\forall\varphi\;\exists n\;\exists\epsilon\;(P_a(n,\epsilon,\varphi(n,\frac{1}{\epsilon}))\leq\frac{1}{2}). \tag{12}$$

Since, by hypothesis, $\{\langle X_n,P_n,F_n\rangle\}_{n\geq 1}$ polynomially converges, fixed $\delta=1/20$, there exists a polynomial ϕ such that

$$\forall\epsilon\;\forall n\;(t\geq\phi(n,\frac{1}{\epsilon})\Rightarrow P_n^{(t)}\{\underline{x}\mid\pi_{F_n}^{(t)}(\underline{x})>\epsilon\}<\frac{1}{20}).$$

From Lemma 4.3 we know that if $t\geq 16/\epsilon^2$ then

$$P_n^{(t)}\{\underline{x}\mid\pi_{F_n}^{(t)}(\underline{x})>\epsilon\}\geq\frac{1}{4}(1-P_a(n,8\epsilon,2t)).$$

If $t\geq\max\{16/\varepsilon^2,\phi(n,1/\varepsilon)\}$, then $\frac{1}{4}(1-P_a(n,8\varepsilon,2t))<\frac{1}{20}$, hence $P_a(n,8\varepsilon,2t)>\frac{4}{5}$.

Fixed a polynomial $\overline{p}(n,1/\varepsilon)$ such that $\frac{1}{2}\overline{p}(n,8/\varepsilon)\geq\max\{16/\varepsilon^2,\phi(n,1/\varepsilon)\}$, we can conclude that

$$\forall\varepsilon\;\forall n\;\left(P_a(n,\varepsilon,\overline{p}(n,\frac{1}{\varepsilon}))>\frac{4}{5}\right). \tag{13}$$

From assertions (12) and (13) follows $\frac{1}{2}>\frac{4}{5}$, a contradiction. □

We are now ready to prove the main result of this section [BeCaMoPa92].

Theorem 4.3 *Given* $\{\langle X_n,P_n,F_n\rangle\}_{n\geq 1}$, *the following conditions are equivalent:*

C1. *The relative frequencies of events in* F_n *converge uniformly polynomially to their probabilities.*

C2. *There exists* $\beta>0$ *such that* $M(n,t)=O(n/t^\beta)$.

C3. *There exists a polynomial* $\psi(n,1/\epsilon)$ *such that* $\forall\epsilon\;\forall n(t\geq\psi(n,1/\epsilon)\Rightarrow M(n,t)\leq\epsilon)$.

Proof.

- C2 ⇒ C3 is readily verified. In fact, condition C2 says there exist α, $\beta>0$ such that $M(n,t)\leq\alpha n/t^\beta$; now, observing that $t\geq(\alpha n/\epsilon)^{\frac{1}{\beta}}$ implies $\alpha n/t^\beta\leq\epsilon$, condition C3 immediately follows.

- C3 $\Rightarrow$ C2. Observe that, as stated by condition C3, there exist $a, b, c > 0$ such that if $t \geq an^b/\epsilon^c$ then $M(n,t) \leq \epsilon$. Solving the first inequality with respect to ϵ gives, in the worst case, $\epsilon = (an^b/t)^{\frac{1}{c}}$, and substituting this ϵ in the second inequality yields $M(n,t) \leq (an^b/t)^{\frac{1}{c}} = a^{\frac{1}{c}} n^{\frac{b}{c}}/t^{\frac{1}{c}}$. If $\frac{b}{c} \leq 1$ we immediately obtain $M(n,t) \leq a^{\frac{1}{c}} n^{\frac{b}{c}}/t^{\frac{1}{c}} \leq a^{\frac{1}{c}} n/t^{\frac{1}{c}}$. Otherwise, if $\frac{b}{c} > 1$, since $M(n,t) \leq 1$, we have $M(n,t) \leq \min\{1, a^{\frac{1}{c}} n^{\frac{b}{c}}/t^{\frac{1}{c}}\} \leq \min\{1, (a^{\frac{1}{c}} n^{\frac{b}{c}}/t^{\frac{1}{c}})^{\frac{c}{b}}\} \leq a^{\frac{1}{b}} n/t^{\frac{1}{b}}$.
- C1 $\Rightarrow$ C3 is an immediate consequence of Theorem 4.2.
- C3 $\Rightarrow$ C1 follows from Theorem 4.1 and from the relation $M(n,t) \leq H_{F_n}(t)/t \leq M(n,t) \cdot \log t$. □

5 Polynomial uniform convergence and polynomial–sample learnability: distribution–dependent case

We will now show that contrarily to what holds for the distribution–free framework, polynomial uniform convergence is a sufficient but not necessary condition for polynomial–sample learnability in the distribution–dependent framework.

Bounds on the sample size sufficient or necessary to learn a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ at approximation ϵ and confidence $1-\delta$ have been given in terms of the minimum cardinality of ϵ-covers [BeIt88]; the following result [Pa91] gives instead an upper bound in terms of the average information per example, $H_{F_n}(t)/t$. We recall that a learning algorithm is said consistent if it always outputs a hypothesis that agrees with the target on every example in the t-sample.

Theorem 5.1 *Given $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, if $\lim_{t\to\infty} H_{F_n}(t)/t = 0$ then for $0 < \epsilon, \delta < 1$, $n \geq 1$ and sample size $t(n, 1/\epsilon, 1/\delta) \geq \frac{128 t_0}{\epsilon^2 \delta}$, where t_0 is such that $H_{F_n}(t_0)/t_0 \leq \epsilon^2/64$, any consistent algorithm for $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is a learning algorithm.*

Proof. Let $f \in F_n$ be the target concept, let $\underline{x}$ be a t-sample and $S_f(\underline{x})$ its corresponding labelled t-sample for f, let $h_{\underline{x}} = A(S_f(\underline{x}))$ be the hypothesis returned by a consistent algorithm A when the sample is $S_f(\underline{x})$. From the definition of consistent algorithm follows that $\nu_{f\triangle h_{\underline{x}}}(\underline{x}) = 0$. Consider the class $f\triangle F_n = \{f\triangle g \mid g \in F_n\}$ of the simmetric differences between f and the functions in F_n. We have

$$P_n^{(t)}\{\underline{x} \mid \mathcal{P}_{f\triangle h_{\underline{x}}} > \epsilon\} = P_n^{(t)}\{\underline{x} \mid |\nu_{f\triangle h_{\underline{x}}}(\underline{x}) - \mathcal{P}_{f\triangle h_{\underline{x}}}| > \epsilon\} \leq P_n^{(t)}\{\underline{x} \mid \pi_{f\triangle F_n}(\underline{x}) > \epsilon\}. \quad (14)$$

Let us estimate the probability of $\{\pi_{f\triangle F_n}(\underline{x}) > \epsilon\}$. We can exploit the relation

$$\Delta_{F_n}(\underline{x}) = \Delta_{f\triangle F_n}(\underline{x}) \quad (15)$$

which follows from the fact that for any $f, g, h \in F_n$ and any t-sample $\underline{x} \in X_n$

$$h \equiv_{\underline{x}} g \leftrightarrow h\triangle f \equiv_{\underline{x}} g\triangle f,$$

where $h \equiv_{\underline{x}} g$ iff $\forall x_i \in \underline{x}\ \ h(x_i) = g(x_i)$. From (15) we obtain that for all $f \in F_n$

$$\frac{H_{F_n}(t)}{t} = \frac{H_{f\triangle F_n}(t)}{t} \quad (16)$$

By hypothesis $\lim_{t\to\infty} H_{F_n}(t)/t = 0$, therefore also $\lim_{t\to\infty} H_{f\triangle F_n}(t)/t = 0$. We can thus apply Lemma 4.2 and, by equation (16), obtain that if $t \geq \frac{128 t_0}{\epsilon^2 \delta}$, where t_0 is such that $H_{F_n}(t_0)/t_0 \leq \epsilon^2/64$, then

$$P_n^{(t)}\{\underline{x} \mid \pi_{f\triangle F_n}(\underline{x}) > \epsilon\} < \delta$$

and therefore, from (14), also

$$P_n\{\underline{x} \mid \mathcal{P}_{f\triangle h_{\underline{x}}} > \epsilon\} < \delta.$$

□

A sufficient condition for the polynomial–sample learnability of a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ in terms of the parameter $M(n,t)$ readily follows from Theorem 4.3, Theorem 5.1, and from the relation $M(n,t) \leq H_{F_n}(t)/t \leq M(n,t)\cdot \log t$. In particular the following holds.

Theorem 5.2 *Given the family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, if there exists $\beta > 0$ such that $M(n,t) = O(n/t^\beta)$ then $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is polynomial–sample learnable.*

Therefore polynomial uniform convergence is a sufficient condition for polynomial–sample learnability. Unexpectedly the converse implication does not hold, contrarily to what happens in the distribution–independent setting. In fact the following theorem holds.

Theorem 5.3 *There exists a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ such that:*

1. *$\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is polynomial–sample learnable.*
2. *$\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is not polynomially uniformly convergent.*

Proof. Let $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ be the family where $X_n = \{0,1\}^n$, P_n is the uniform distribution over X_n and $F_n = \{f | f^{-1}(1) \subseteq \{0,1\}^n \wedge \# f^{-1}(1) \leq 2^{\frac{n}{2}}\}$.

1. $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is polynomial–sample learnable. In fact, consider the following algorithm.

   Algorithm $\mathcal{A}$

   - *input*: n, ϵ, δ
   - **if** $\epsilon \geq 2^{-\frac{n}{2}}$
     then output as hypothesis the function $\emptyset(x) = 0$ for all x (without making calls to the oracle)
     else (i.e. $\epsilon < 2^{-\frac{n}{2}}$) make $t = (\frac{4}{\epsilon}\log\frac{2}{\delta} + \frac{8}{\epsilon^2}\log\frac{13}{\epsilon})$ calls to the oracle to obtain a t-sample $(\langle x_1, f(x_1)\rangle, \ldots, \langle x_t, f(x_t)\rangle)$; output as hypothesis the set $\{x_i \mid f(x_i) = 1\}$

   Algorithm $\mathcal{A}$ is a learning algorithm for $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$. Let $f \in F_n$ be the function to be learned.
   If $\epsilon \geq 2^{-\frac{n}{2}}$ then, denoting by $\oplus$ the boolean function exclusive or,

   $$P_n\{x | (f \oplus \emptyset)(x) = 1\} = \frac{\# f^{-1}(1)}{2^n} \leq 2^{-\frac{n}{2}} \leq \epsilon.$$

If $\epsilon < 2^{-\frac{n}{2}}$, since $\mathcal{A}$ is a consistent algorithm, it needs $(\frac{4}{\epsilon}\log\frac{2}{\delta} + \frac{8\mathrm{VCdim}(F_n)}{\epsilon}\log\frac{13}{\epsilon})$ examples to learn F_n at approximation ϵ and confidence $1-\delta$ [BlEhHaWa89]. Observing that $\mathrm{VCdim}(F_n) = 2^{n/2} < 1/\epsilon$, then $(\frac{4}{\epsilon}\log\frac{2}{\delta} + \frac{8}{\epsilon^2}\log\frac{13}{\epsilon})$ examples are sufficient.

2. $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is not polynomially uniformly convergent. In fact, if the sample size satisfies the constraint $t \leq 2^{n/2}$ then, denoting $\{x_1, \ldots, x_t\}$ by s,

$$\pi^{(t)}_{F_n}(x_1, \ldots, x_t) = \sup_{f\in F_n} |\nu_f(x_1, \ldots, x_t) - \mathcal{P}_f| \geq |\nu_s(x_1, \ldots, x_t) - \mathcal{P}_s| \geq 1 - \frac{t}{2^n} > \frac{1}{4};$$

therefore $t \leq 2^{\frac{n}{2}} \Rightarrow P^{(t)}_n\{\underline{x} \mid \pi^{(t)}_{F_n}(\underline{x}) > 1/4\} = 1$. This implies that the sample size $t(n, 1/\epsilon, 1/\delta)$ necessary for the uniform convergence at approximation ϵ and confidence $1-\delta$, must satisfy, for $\epsilon = 1/4$ and $\delta = 1/2$, $t(n, 4, 2) \geq 2^{n/2}$; therefore $t(n, 1/\epsilon, 1/\delta)$ can not be bounded by a polynomial in $n, 1/\epsilon, 1/\delta$. □

6 Conclusions and open problems

In this paper we have investigated the problem of polynomial–sample learnability in relation to the concept of polynomial uniform convergence. Even though the results are stated for the boolean domain $\{0,1\}^n$, they also hold for infinite domains such as E^n, provided that certain measurability constraints on the concept class (see [BlEhHaWa89]) are satisfied.

In the distribution–dependent context we have characterized the property of polynomial uniform convergence of $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ by means of the parameter $M(n,t)$. In particular we proved that $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$ is polynomially uniformly convergent iff there exists $\beta > 0$ such that $M(n,t) = O(n/t^\beta)$. No attempt has been made to obtain better upper and lower bounds on the sample size in terms of $M(n,t)$; therefore it is an open problem whether there are tight bounds on the sample size in terms of this parameter.

With respect to the relationship between polynomial uniform convergence and PAC learning in the distribution–dependent context, we have shown that if $M(n,t) = O(n/t^\beta)$ for a family $\{\langle X_n, P_n, F_n\rangle\}_{n\geq 1}$, then the family can be PAC learned with a sample of size bounded by a polynomial in n, $1/\epsilon$, $1/\delta$. We have also shown that the converse implication does not hold. Therefore in the distribution–dependent context it still is an open problem to find an "informational" parameter characterizing the polynomial–sample learnability.

Acknowledgements

We thank Bruno Codenotti and Alberto Marchetti–Spaccamela for reading preliminary versions of this paper.

References

[BeIt88] G. Benedek, A. Itai. "Learnability by Fixed Distributions". *Proc. 1988 Workshop on Computational Learning Theory* (1988) 80-90.

[BeBeMa89] S. Ben–David, G. Benedek, Y. Mansour. "A Parametrization Scheme for Classifying Models of Learnability". *Proc. 1989 Workshop on Computational Learning Theory* (1989) 285-302.

[BeCaMoPa92] A. Bertoni, P. Campadelli, A. Morpurgo, S. Panizza. "Polynomial Uniform Convergence of Relative Frequencies to Probabilities". *Advances in Neural Information Processing Systems 4*, San Mateo, CA: Morgan Kaufmann Publishers (1992) 904–911.

[BlEhHaWa89] A. Blumer, A. Ehrenfeucht, D. Haussler, K. Warmuth. "Learnability and the Vapnik–Chervonenkis Dimension". *J. ACM* **36** (1989) 929–965.

[Na88] B. Natarajan. "Learning over Classes of Distributions". *Proc. 1988 Workshop on Computational Learning Theory* (1988) 408–409.

[Na92] B. Natarajan. "Probably Approximate Learning Over Classes of Distributions". *Siam J. Comput.* **21** (3) (1992).

[Pa91] S. Panizza. "Apprendimento PAC con distribuzione di probabilità fissata". Tesi di Laurea, Università degli studi di Milano, Dipartimento di Scienze dell'Informazione, A.A. 1990-91 (1991).

[Sa72] N. Sauer. "On the Density of Families of Sets". *J. Combinatorial Theory* (A) **13** (1972) 145–147.

[Va84] L.G. Valiant. "A Theory of the Learnable". *Communications of the ACM* **27** (1984) 1134–1142.

[VaCh71] V.N. Vapnik, A.Ya. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". *Theory of Prob. and its Appl.* **16** (2) (1971) 265–280.