JMLR: Workshop and Conference Proceedings vol 40:1–12, 2015

An Almost Optimal PAC Algorithm

Hans Ulrich Simon HANS.SIMON@RUB.DE
Horst Görtz Institute for IT Security and Faculty of Mathematics, Ruhr-University Bochum, Germany

Abstract

The best currently known general lower and upper bounds on the number of labeled examples needed for learning a concept class in the PAC framework (the realizable case) do not perfectly match: they leave a gap of order $\log(1/\epsilon)$ (resp. a gap which is logarithmic in another one of the relevant parameters). It is an unresolved question whether there exists an "optimal PAC algorithm" which establishes a general upper bound with precisely the same order of magnitude as the general lower bound. According to a result of Auer and Ortner (2007), there is no way for showing that arbitrary consistent algorithms are optimal because they can provably differ from optimality by factor $\log(1/\epsilon)$. In contrast to this result, we show that every consistent algorithm L (even a provably suboptimal one) induces a family $(L_K)_{K\geq 1}$ of PAC algorithms (with $2K-1$ calls of L as a subroutine) which come very close to optimality: the number of labeled examples needed by L_K exceeds the general lower bound only by factor $\ell_K(1/\epsilon)$ where ℓ_K denotes (a truncated version of) the K-times iterated logarithm. Moreover, L_K is applicable to any concept class C of finite VC-dimension and it can be implemented efficiently whenever the consistency problem for C is feasible. We show furthermore that, for every consistent algorithm L, L_2 is an optimal PAC algorithm for precisely the same concept classes which were used by Auer and Ortner (2007) for showing the existence of suboptimal consistent algorithms. This can be seen as an indication that L_K may have an even better performance than it is suggested by our worstcase analysis.

Keywords: PAC-learning, estimation error, sample size, optimal PAC algorithm, majority vote

1. Introduction

More than thirty years after the introduction of the PAC-learning framework (the realizable case) by Valiant (1984), it is still not completely known how many labeled examples are needed for learning successfully in this model (i.e., for returning a hypothesis that, with a probability of at least $1-\delta$, is correct up to an error of at most ϵ). The following lower and upper bounds in terms of ϵ, δ and the VC-dimension d of the underlying concept class are known:

- Ehrenfeucht et al. (1989) have shown that every PAC algorithm needs at least $\Omega\left(\frac{1}{\epsilon}\left(d+\log\left(\frac{1}{\delta}\right)\right)\right)$ labeled examples.
- Blumer et al. (1989) have shown that every consistent algorithm needs no more than $O\left(\frac{1}{\epsilon}\left(d\log\left(\frac{1}{\epsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right)$ labeled examples. As shown by Auer and Ortner (2007), this upper bound is actually tight for some consistent algorithms.

These bounds coincide (in their order of magnitude) except for the factor $\log(1/\epsilon)$ that occurs in the upper bound only. Warmuth (2004) raised the question whether there exists an "optimal PAC algorithm" which does not need more than $O\left(\frac{1}{\epsilon}\left(d+\log\left(\frac{1}{\delta}\right)\right)\right)$ labeled examples (the same order of magnitude as in the general lower bound). He conjectured that the so-called "1-inclusion

graph algorithm" of Haussler et al. (1994) fits this purpose. While this conjecture is still open, the following progress has been made:

- Hanneke (2009) has shown that the additional factor $\log(1/\epsilon)$ in the upper bound can be replaced by the logarithm of the so-called disagreement coefficient associated with C. Note that a constant disagreement coefficient would blur the distinction between the upper and the lower bound.

- Building on the work by Hanneke (2009) but using an alternative definition[1] of the disagreement coefficient, Darnstädt (2014) proved that the closure algorithm (which always returns the smallest consistent hypothesis from C) is an optimal PAC algorithm for intersection-closed classes.[2] In view of Warmuth's conjecture, it is interesting to note that the 1-inclusion graph algorithm, when applied to intersection-closed classes, collapses to the closure algorithm.

In this paper, we present a family $(L_K)_{K\geq 1}$ of PAC algorithms which are almost optimal in the following precise sense. For $z > 0$, let $\log^{(1)}(z) = \log(z)$ and $\log^{(K)}(z) = \log(\log^{(K-1)}(z))$, i.e., $\log^{(K)}$ is the K-times iterated logarithm. Let $\ell_K(z) = \max\{2, \log^{(K)}(z)\}$ be a truncated version of $\log^{(K)}$ whose values cannot drop below 2. Then L_K applied to a concept class of VC-dimension d requires only $O\left(\frac{1}{\epsilon}\left(d\ell_K\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ labeled examples for PAC-learning C. Compared to the general lower bound, this leaves a gap of size $\ell_K(1/\epsilon)$ only. Noting that 10^{80} is considered as an upper bound on the number of atoms in the universe and noting that $\log^{(4)}(10^{80}) < 1.6$ (so that $\ell_4(10^{80}) = 2$), it becomes evident that $\log^{(K)}$ and ℓ_K are functions which go to infinity at an extremely slow rate (and this already holds for quite moderate values of K). While L_K is always optimal up to factor $\ell_K(1/\epsilon)$, it is occasionally even better: we show that L_2 is an optimal PAC algorithm for the same concept classes which were used by Auer and Ortner (2007) for showing the existence of suboptimal consistent algorithms.

The technique that we will use for analyzing L_K is related to a technique used by Hanneke (2009) for analyzing consistent algorithms. Hanneke brought the disagreement coefficient into play by decomposing a given sequence S of labeled examples into subsequences $S_1, S_2, \ldots$ and by analyzing the hypotheses from C which are consistent with S_k (the k-th subsequence) not against the (unconditioned) domain distribution P but against P conditioned to hitting the so-called disagreement region induced by $S_1 \cup \ldots \cup S_{k-1}$. The reader familiar with Hanneke's work may detect some similarities to it in our approach but we would also like to stress the following differences:

- Our algorithm L_K actually performs the decomposition of S into $2K-1$ subsequences and creates one individual hypothesis h_k from C for every individual subsequence S_k (while in Hanneke's work the decomposition of S happens in the analysis only and has no influence to what the algorithm does).

- Instead of returning a single hypothesis from C that is consistent with S (as Hanneke does), we return the majority vote over the individual hypotheses $h_1, \ldots, h_{2K-1}$.

- When analyzing the hypotheses from C which are consistent with S_k, we condition P not on hitting a disagreement region (as Hanneke does) but on hitting the common error region of some of the hypotheses from $\{h_1, \ldots, h_{k-1}\}$.

1. originally suggested by Darnstädt et al. (2014)
2. For a proper subclass of intersection-closed classes, this had been shown before by Auer and Ortner (2007).

The remainder of the paper is organized as follows. Section 2 provides the reader with the relevant formal definitions, notations and facts. In Section 3, we state and prove the main results. In Section 4, we show that L_2 is optimal for a concept class for which L_1 fails to be optimal by factor $\log(1/\epsilon)$. In the final Section 5, we discuss some algorithmic modifications of the family $(L_K)_{K\geq 1}$ and some efficiency issues. Moreover, we raise some open questions.

2. Definitions, Notations and Facts

Throughout this paper, $\log$ (resp. $\ln$) denotes the logarithm to base 2 (resp. base e). We briefly note that the iterated logarithm and its truncated version (as defined in the introduction) are related as follows:

$$\log^{(k)}(z) = \log(\log^{(k-1)}(z)) \leq \log(\ell_{k-1}(z)) \leq \ell_k(z) \ . \tag{1}$$

Let Z be a random variable that is binomially distributed with parameters p, m. According to one of the bounds by Chernoff (1952), we have that

$$\Pr\left[Z < \frac{1}{2}mp\right] \leq e^{-mp/8} \ . \tag{2}$$

We assume the reader to be familiar with the PAC-learning model in the realizable case but call briefly into mind some of the central definitions and thereby fix some notation.

Let X be a set and let C, H be families of functions from X to $\{0,1\}$. The elements of X (resp. of $X \times \{0,1\}$) are called *examples* (resp. *labeled examples*). A sequence $S \in (X \times \{0,1\})^*$ of labeled examples is called a *sample*. For $x \in X^m$ and $f \in C$, we define $S_f(x) = [(x_1, f(x_1)), \ldots, (x_m, f(x_m))]$. The elements of C (resp. H) are called *concepts* (resp. *hypotheses*). Let P be a probability measure on X and let $f \in C$. The error of a hypothesis h w.r.t. P and f is defined as follows:

$$er_{P,f}(h) = P(\{x \in X : \ h(x) \neq f(x)\}) \ .$$

Let $L : (X \times \{0,1\})^* \to H$ be a function that maps a sample S to a hypothesis $L(S) \in H$. We say that L *is a PAC function* for the concept class C if there exists a function $m(\epsilon, \delta)$ such that the following holds. For each concept $f \in C$, for each probability measure P on X, for each choice of $0 < \epsilon, \delta < 1$ and for all $m \geq m(\epsilon, \delta)$, we have that

$$P^m(\{x \in X^m : \ er_{P,f}(L(S_f(x))) \leq \epsilon\}) \geq 1 - \delta \ . \tag{3}$$

An algorithm that computes a PAC function is simply called a *PAC algorithm*. In the context of Condition (3), f is called the *target concept*, ϵ is called the *accuracy parameter*, and δ is called the *confidence parameter*. The *sample complexity function* $m_L(\epsilon, \delta)$ of L is defined as follows. For each $0 < \epsilon, \delta < 1$, $m_L(\epsilon, \delta)$ is the smallest number such that (3) holds for all choices of $f \in C$ and P provided that $m \geq m_L(\epsilon, \delta)$. The *estimation error* $\epsilon_L(m, \delta)$ of L is defined analogously: for each $m \geq 1$ and each $\delta \in (0,1)$, $\epsilon_L(m, \delta)$ is the smallest real in the interval $(0,1)$ such that (3) holds for all choices of $f \in C$ and P provided that $\epsilon \geq \epsilon_L(m, \delta)$. A hypothesis $h \in H$ is said to be *consistent* with $S = [(x_1, b_1), \ldots, (x_m, b_m)] \in (X \times \{0,1\})^m$ if $h(x_i) = b_i$ for $i = 1, \ldots, m$. The set of all hypotheses in H that are consistent with S is called the *version space* for S in H and is denoted as $V_H(S)$. We say that $L : (X \times \{0,1\})^* \to C$ is a *consistent function* if, for every $f \in C$ and every $x \in X^*$, we have that $L(S_f(x)) \in V_C(S_f(x))$. An algorithm that computes a consistent function is said to be a *consistent algorithm*. For finite concept classes C, the following is known:

Theorem 1 (Blumer et al. (1987)) *Let C be a finite concept class over X. Then the following holds for every $f \in C$, every probability measure P on X and every choice of $m \geq 1, 0 < \epsilon, \delta < 1$. With a probability of at least $1 - \delta$ (taken over $x \sim P^m$), every $h \in V_C(S_f(x))$ satisfies*

$$er_{f,P}(h) \leq \frac{1}{m} \ln\left(\frac{|C|}{\delta}\right) \quad .$$

Thus, every consistent function $L : (X \times \{0,1\})^* \to C$ is a PAC function for C whose estimation error and sample complexity function are bounded from above as follows:

$$\epsilon_L(m,\delta) \leq \frac{1}{m} \ln\left(\frac{|C|}{\delta}\right) \quad \textit{and} \quad m_L(\epsilon,\delta) \leq \frac{1}{\epsilon} \ln\left(\frac{|C|}{\delta}\right) \quad .$$

A sequence $x \in X^m$ is said to be *shattered* by C if, for every $b \in \{0,1\}^m$, we have that $V_C([(x_1,b_1),\ldots,(x_m,b_m)]) \neq \emptyset$. The *VC-dimension* of C is defined as the largest m such that X^m contains a sequence which is shattered by C (resp. as ∞ if shattered sequences can become arbitrarily long).

Theorem 2 (Blumer et al. (1989)) *Let C be a concept class over X and let d denote its VC-dimension. Then the following holds for every $f \in C$, every probability measure P on X and every choice of $m \geq d, 0 < \epsilon, \delta < 1$. With a probability of at least $1 - \delta$ (taken over $x \sim P^m$), every $h \in V_C(S_f(x))$ satisfies $er_{f,P}(h) \leq \epsilon'_{ub}(m,d,\delta)$ where*

$$\epsilon'_{ub}(m,d,\delta) = \frac{2}{m}\left(d \log\left(\frac{2em}{d}\right) + \log\left(\frac{2}{\delta}\right)\right) \quad .$$

Thus, every consistent function $L : (X \times \{0,1\})^* \to C$ is a PAC function for C whose estimation error satisfies $\epsilon_L(m,\delta) \leq \epsilon'_{ub}(m,d,\delta)$. Moreover, $\epsilon_L(m,\delta) \leq \epsilon'_{ub}(m,d,\delta)$ implies that

$$m_L(\epsilon,\delta) \leq \max\left\{\frac{8d}{\epsilon}\log\left(\frac{13}{\epsilon}\right), \frac{4}{\epsilon}\log\left(\frac{2}{\delta}\right)\right\} \quad .$$

In this paper, it will be technically more convenient to use the following bound instead of ϵ'_{ub}:

$$\epsilon_{ub}(m,d,\delta) = \frac{4}{m} \cdot \max\left\{d \log\left(\frac{2em}{d}\right), \log\left(\frac{2}{\delta}\right)\right\} \quad . \tag{4}$$

Note that $\epsilon'_{ub}(m,d,\delta) \leq \epsilon_{ub}(m,d,\delta) \leq 2\epsilon'_{ub}(m,d,\delta)$. for each choice of (m,d,δ), i.e., the upper bound ϵ_{ub} is (slightly) weaker than ϵ'_{ub}.

3. Almost Optimality of the Majority Vote

Let C be a concept class over domain X and let $L : (X \times \{0,1\})^* \to C$ be a consistent function. For each $K \geq 1$, we define the function L_K algorithmically as follows:

1. Given a sample $S \in (X \times \{0,1\})^{(2K-1)m}$, decompose S into $2K-1$ subsamples $S_1,\ldots,S_{2K-1}$ of size m, respectively, so that $S = (S_1,\ldots,S_{2K-1})$.

2. For $k = 1,\ldots,2K-1$, let $h_k = L(S_1 \cup \ldots \cup S_k)$.

3. Let $L_K(S)$ be the majority vote over $h_1, \ldots, h_{2K-1}$, i.e., $L_K(S)$ assigns the label 1 to $x \in X$ iff at least K of the hypotheses $h_1, \ldots, h_{2K-1}$ make the same assignment.

In this description of L_K, we assumed that the sample size, $|S|$, is a multiple of $2K-1$. This assumption can of course be removed by splitting a sample of any size into $2K-1$ subsamples of almost equal size. But in the sequel, we maintain the assumption because it will allow for an easier exposition of the central arguments. Here comes the main result of this section:

Theorem 3 *Let C be a concept class of VC-dimension d over domain X. Let $L : (X \times \{0,1\})^* \to C$ be any consistent function. Then, for every constant $K \geq 1$, L_K is a PAC function for C whose estimation error is bounded from above as follows:*[3]

$$\epsilon_{L_K}(m,\delta) = O\left(\frac{1}{m}\left(d\ell_K\left(\frac{m}{d}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) \ .$$

Moreover, this implies that the sample complexity function of L_K is bounded from above as follows:

$$m_{L_K}(\epsilon,\delta) = O\left(\frac{1}{\epsilon}\left(d\ell_K\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) \ .$$

While the bound $O\left(\frac{1}{\epsilon}\left(d\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ on m_L differs from optimality by factor $\log(1/\epsilon)$, the above bound on m_{L_K} differs from optimality only by factor $\ell_K(1/\epsilon)$. Note that $L_1 = L$. In other words: for $K = 1$, the function L_K collapses to the consistent function L. In view of the results of Auer and Ortner (2007) that we already mentioned in the introduction, this implies that the above upper bound on $m_{L_1}(\epsilon,\delta)$ is tight for some choices of L.

The remainder of this section is devoted to the analysis of the function L_K. We start with Lemma 4 which provides an upper bound on the members of a recursively defined sequence $(\epsilon_k)_{k\geq 1}$. When the lemma is applied later within the proof of Theorem 5, ϵ_k will be the total probability mass of a common error region of a fixed selection of k hypotheses (chosen from $h_1, \ldots, h_{2K-1}$), and this will help us to upper-bound the estimation error of L_K. The proof of Lemma 4 is somewhat tedious but instructive in so far as it demonstrates how the iterated logarithm comes into play. However, the reader not willing to dip into the technical details of the proof may simply skip it without loss of continuity.

Lemma 4 *Let $(\epsilon_k)_{k\geq 1}$ be a sequence of functions in m, d, δ that satisfies the recursion*

$$\epsilon_1 = \epsilon_{ub}(m,d,\delta) \ \text{ and } \ \epsilon_k = \epsilon_{k-1} \cdot \min\left\{1, \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, d, \delta\right)\right\} \ \text{ for } k \geq 2 \ .$$

Let $(c_k)_{k\geq 1}$ be the sequence that is recursively given by

$$c_1 = 0 \ , \ c_2 = \log(4e) \ \text{ and } \ c_k = \log(8e) + \log(c_{k-1}) \ \text{ for } k \geq 3 \ .$$

Let $a_1 = 4$ and $a_k = 8$ for $k \geq 2$, and let finally

$$\epsilon_{ub}^{(k)}(m,d,\delta) = \frac{a_k}{m} \cdot \max\left\{c_k d + d\ell_k\left(\frac{2em}{d}\right), \log\left(\frac{2}{\delta}\right)\right\} \ . \tag{5}$$

3. The Big-O notation hides only universal constants or constant factors depending on K.

Then $(c_k)_{k\geq 1}$ is a strictly increasing sequence that converges to the unique number $c > 3$ which satisfies $c = \log(8e) + \log(c)$.[4] *Moreover, for all $m, d \geq 1$, all $0 < \delta < 1$ and for all $k \geq 1$, we have that $\epsilon_k \leq \epsilon_{ub}^{(k)}(m, d, \delta)$.*

Proof The assertion on the sequence $(c_k)_{k\geq 1}$ is obvious. A comparison of (5) and (4) in combination with $\log(z) \leq \ell_1(z)$ shows that $\epsilon_{ub}(m, d, \delta) \leq \epsilon_{ub}^{(1)}(m, d, \delta)$. Thus $\epsilon_1 \leq \epsilon_{ub}^{(1)}(m, d, \delta)$, as desired. We may therefore assume inductively that

$$\epsilon_{k-1} \leq \epsilon_{ub}^{(k-1)}(m, d, \delta) = \frac{a_{k-1}}{m} \cdot \max\left\{c_{k-1}d + d\ell_{k-1}\left(\frac{2em}{d}\right), \log\left(\frac{2}{\delta}\right)\right\} \quad .$$

If the maximum equals $\log(2/\delta)$, then $\epsilon_{k-1} \leq \epsilon_{ub}^{(k-1)}(m, d, \delta) = \frac{a_{k-1}}{m}\log(2/\delta)$, and this upper bound also applies to $\epsilon_k \leq \epsilon_{k-1}$. Thus, let us assume that the maximum equals the other term so that

$$\epsilon_{k-1} \leq \frac{a_{k-1}d}{m} \cdot \left(c_{k-1} + \ell_{k-1}\left(\frac{2em}{d}\right)\right) \quad . \tag{6}$$

The recursive definition of ϵ_k in the lemma and an application of (4) with $\frac{1}{2}\epsilon_{k-1}m$ at the place of m yields the following:

$$\epsilon_k \leq \epsilon_{k-1} \cdot \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, d, \delta\right) = \frac{8}{m} \cdot \max\left\{d\log\left(\frac{e\epsilon_{k-1}m}{d}\right), \log\left(\frac{2}{\delta}\right)\right\} \quad .$$

Again we will be done if the maximum equals $\log(2/\delta)$. We may therefore assume that

$$\epsilon_k \leq \frac{8d}{m} \cdot \log\left(\frac{e\epsilon_{k-1}m}{d}\right) \quad . \tag{7}$$

Recall that $c_1 = 0$ and $a_1 = 4$. An application of (6) for $k = 2$ yields

$$\frac{e\epsilon_1 m}{d} \leq 4e \cdot \ell_1\left(\frac{2em}{d}\right) \quad .$$

We may continue with the calculation (7) and conclude that

$$\epsilon_2 \leq \frac{8d}{m} \cdot \log\left(4e \cdot \ell_1\left(\frac{2em}{d}\right)\right) \overset{(1)}{\leq} \frac{8d}{m} \cdot \left(\log(4e) + \ell_2\left(\frac{2em}{d}\right)\right) \quad .$$

Since $c_2 = \log(4e)$, we may conclude that $\epsilon_2 \leq \epsilon_{ub}^{(2)}(m, d, \delta)$. Recall that $a_k = 8$ for $k \geq 2$ and note that $c_k > 2$ for $k \geq 2$. An application of (6) for $k \geq 3$ yields

$$\frac{e\epsilon_{k-1}m}{d} \leq 8e \cdot \left(c_{k-1} + \ell_{k-1}\left(\frac{2em}{d}\right)\right) \leq 8e \cdot c_{k-1} \cdot \ell_{k-1}\left(\frac{2em}{d}\right)$$

where the latter inequality holds because $c_{k-1}, \ell_{k-1}(2em/d) \geq 2$ (and is based on the fact that $a + b \leq ab$ provided that $a, b \geq 2$). Again we may continue with the calculation (7) and conclude that

$$\epsilon_k \leq \frac{8d}{m} \cdot \log\left(8e \cdot c_{k-1} \cdot \ell_{k-1}\left(\frac{2em}{d}\right)\right) \overset{(1)}{\leq} \frac{8d}{m} \cdot \left(\log(8e) + \log(c_{k-1}) + \ell_k\left(\frac{2em}{d}\right)\right) \quad .$$

4. A numerical computation shows that c is smaller than (and approximately equal to) 7.35.

Since $\log(8e) + \log(c_{k-1}) = c_k$ for $k \geq 3$, we may conclude that $\epsilon_k \leq \epsilon_{ub}^{(k)}(m, d, \delta)$, as desired. ■

We are now prepared for the proof of Theorem 3 (our main result). We will actually prove the following stronger result, which does not hide the constant factors depending on K:

Theorem 5 *Let C be a concept class over X and let d denote its VC-dimension. Let $L : (X \times \{0,1\})^* \to C$ be any consistent function. Then, for every $K \geq 2$, L_K is a PAC function for C whose estimation error satisfies*

$$\epsilon_{L_K}\left((2K-1)m, \binom{2K-1}{K}(2K-1)\delta\right) \leq \binom{2K-1}{K}\epsilon_{ub}^{(K)}(m, d, \delta) \ . \tag{8}$$

Proof Let $S \in (X \times \{0,1\})^{(2K-1)m}$ denote the sample whose instances are drawn at random according to $P^{(2K-1)m}$ and labeled according to the target concept $f \in C$. Let $S_1, \ldots, S_{2K-1}$ denote the corresponding subsamples of size m, respectively. Let $h = L_K(S)$, i.e., h is the majority vote over the hypotheses $h_1, \ldots, h_{2K-1}$ where $h_k = L(S_1 \cup \ldots \cup S_k)$ for $k = 1, \ldots, 2K-1$. Let $E = \{x \in X : h(x) \neq f(x)\}$. For $j = 1, \ldots, 2K-1$, let $E_j = \{x \in X : h_j(x) \neq f(x)\}$. For every $J \subseteq \{1, \ldots, 2K-1\}$, let $E_J = \cap_{j \in J} E_j$. Let M denote the family of subsets of $\{1, \ldots, 2K-1\}$ that have cardinality K. With these notations, we clearly obtain

$$E = \bigcup_{J \in M} E_J = \bigcup_{J \in M} \bigcap_{j \in J} E_j \ .$$

From the union bound, we get

$$P(E) \leq \sum_{J \in M} P(E_J) \leq \binom{2K-1}{K} \max_{J \in M} P(E_J) \ .$$

Our goal is to show the following for each individual set $J \in M$: with a probability of at least $(2K-1)\delta$ (taken over the random sample S), we have that $P(E_J) \leq \epsilon_{ub}^{(K)}(m, d, \delta)$. If this goal can be achieved, the following conclusion will be obtained immediately: with a probability of at least $1 - \binom{2K-1}{K}(2K-1)\delta$, we have that $P(E) \leq \binom{2K-1}{K} \cdot \epsilon_{ub}^{(K)}(m, d, \delta)$. Since this confirms the bound (8), all that remains to be done is showing that the above assertion on an individual set $J \in M$ is true. To this end, let $J = \{j(1), \ldots, j(K)\}$ with $1 \leq j(1) < \ldots < j(K) \leq 2K-1$ be a fixed, but otherwise arbitrary, set in M. For sake of brevity, we set

$$E'_k = \bigcap_{l=1}^{k} E_{j(l)} \ , \ \epsilon_k = P(E'_k) \ \text{ and } \ S'_k = \{(x, b) \in S_{j(k)} : \ x \in E'_{k-1}\}$$

for $k = 1, \ldots, K$ with the convention that $\epsilon_0 = 1$ and $E'_0 = X$ so that $S'_1 = S_{j(1)}$. Note that, with this notation, $E'_K = E_J$. Obviously, $E'_{k'} \subseteq E'_k$ for $k' \geq k$. Specifically, $E_J = E'_K \subseteq E'_k$ for $k = 1, \ldots, K$. Thus,

$$\forall k = 1, \ldots, K : P(E_J) \leq P(E'_k) = \epsilon_k \ . \tag{9}$$

From the theorem on compound probabilities, we obtain

$$\epsilon_k = P(E_{j(k)} | E'_{k-1}) \cdot \epsilon_{k-1} = \prod_{l=1}^{k} P(E_{j(l)} | E'_{l-1}) \tag{10}$$

for $k = 1, \ldots, K$. Specifically,

$$P(E_J) = \epsilon_K = \prod_{k=1}^{K} P(E_{j(k)}|E'_{k-1}) \ . \tag{11}$$

The equation (11) suggests to evaluate the hypothesis $h_{j(k)}$ against the conditional probability $P(\cdot|E'_{k-1})$. Since we have defined S'_k as the subsample of $S_{j(k)}$ whose instances fall into E'_{k-1}, it follows that S'_k is a sample whose instances are independently drawn at random according to $P(\cdot|E'_{k-1})$. For this reason, we call $m_k = |S'_k|$ the *effective sample size* at stage k. Note that, given E'_{k-1}, m_k is binomially distributed with parameters ϵ_{k-1}, m. Moreover, $m_1 = |S'_1 \cap X| = m$, i.e. the effective sample size at stage 1 equals m. The following events, with k ranging from 2 to K, bear the danger of making $P(E_J)$ large:

$$\begin{aligned}
B_1 &\Leftrightarrow \exists g \in V_C(S'_1) : P[g(x) \neq f(x)] > \epsilon_{ub}(m, d, \delta) \ . \\
B'_k &\Leftrightarrow \left(\epsilon_{k-1} \geq \frac{8\ln(1/\delta)}{m}\right) \wedge \left(m_k < \frac{1}{2}\epsilon_{k-1}m\right) \ . \\
B_k &\Leftrightarrow \left(m_k \geq \frac{1}{2}\epsilon_{k-1}m\right) \\
&\quad \wedge \left(\exists g \in V_C(S'_k) : P[g(x) \neq f(x)|x \in E'_{k-1}] > \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, d, \delta\right)\right) \ .
\end{aligned}$$

We make the following observations concerning the probabilities of B_k and B'_k (taken over the random sample S):

- The probability of B_1 is bounded by δ according to Theorem 2.
- The probability of B'_k is bounded by δ according to the Chernoff bound (2) with m_k in the role of Z and ϵ_{k-1} in the role of p.
- The probability of B_k is bounded by δ according to Theorem 2 with $P(\cdot|E'_{k-1})$ in the role of P.

We may conclude that the probability of the event $B = \cup_{k=1}^{K} B_k \cup \cup_{k=2}^{K} B'_k$ is bounded by $(2K-1)\delta$. The following claim is the final piece of puzzle in our proof:

Claim: If the event B does not occur, then $P(E_J) \leq \epsilon_{ub}^{(K)}(m, d, \delta)$.

Assume that B does not occur. If $\epsilon_{k-1} < \frac{8\ln(1/\delta)}{m}$ for some $k \in \{2, \ldots, K\}$, then $P(E_J) < \frac{8\ln(1/\delta)}{m}$ according to (9), and we are done. Thus we may safely assume that $\epsilon_{k-1} \geq \frac{8\ln(1/\delta)}{m}$ for $k = 2, \ldots, K$. Since, by assumption, none of the events B'_k occurs, we may conclude that $m_k \geq \frac{1}{2}\epsilon_{k-1}m$ for $k = 2, \ldots, K$. Since none of the events B_k occurs, we may furthermore conclude that, for $k = 2, \ldots, K$, the following holds:

$$\forall g \in V_C(S'_k) : P(g(x) \neq f(x)|x \in E'_{k-1}) \leq \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, \delta\right) \ .$$

Moreover,

$$\forall g \in V_C(S_1) : P(g(x) \neq f(x)) \leq \epsilon_{ub}(m, d, \delta) \ .$$

Since $h_{j(k)}$ is consistent with $S_1 \cup \ldots \cup S_{j(k)}$, we clearly have $h_{j(k)} \in V_C(S'_k)$ resp. $h_{j(1)} \in V_C(S_1)$.[5] Thus,

$$P(E_{j(k)}|E'_{k-1}) \leq \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, d, \delta\right) \tag{12}$$

holds for $k = 2, \ldots, K$. Moreover, $\epsilon_1 = P(E_{j(1)}) \leq \epsilon_{ub}(m, d, \delta)$. We arrive at the following recursion on $\epsilon_k = P(E'_k)$:

$$\epsilon_1 \leq \epsilon_{ub}(m, d, \delta) \text{ and } \epsilon_k \overset{(10),(12)}{\leq} \epsilon_{k-1} \cdot \epsilon_{ub}\left(\frac{1}{2}\epsilon_{k-1}m, d, \delta\right) .$$

Thus the sequence $(\epsilon_k)_{k=1,\ldots,K}$ satisfies the assumptions in Lemma 4. Since $P(E_J) = \epsilon_K$, the claim is now immediate from this lemma. ■

The following upper bounds on $\epsilon_{L_K}(m, \delta)$ and on $m_{L_K}(m, \delta)$ are obtained from (8) and from the fact that

$$\binom{2K-1}{K} = O\left(\frac{2^{2K}}{\sqrt{K}}\right)$$

by a straightforward calculation:

$$\epsilon_{L_K}(m, \delta) = O\left(\frac{2^{2K}\sqrt{K}}{m}\left(d\ell_K\left(\frac{m}{d}\right) + K + \log\left(\frac{1}{\delta}\right)\right)\right) . \tag{13}$$

$$m_{L_K}(m, \delta) = O\left(\frac{2^{2K}\sqrt{K}}{\epsilon}\left(d\ell_K\left(\frac{1}{\epsilon}\right) + K + \log\left(\frac{1}{\delta}\right)\right)\right) . \tag{14}$$

Theorem 3 results from these bounds simply by hiding the constant factors depending on K.

4. Reaching Optimality by Going with the Majority

We start with a result by Auer and Ortner (2007) which shows that consistent algorithms (and therefore L_K with $K = 1$) can fail to be optimal by factor $\log(1/\epsilon)$:

Theorem 6 (Auer and Ortner (2007)) *Let X be an infinite domain and, for every $d \geq 1$, let C_d be the class of concepts over X that assign the label 1 to at most d elements of X (a class which is known to have VC-dimension d). Let $L : (X \times \{0,1\})^* \to C_d$ be a consistent function that returns a largest consistent concept. Then $m_L(\epsilon, 1/2) = \Omega\left(\frac{d}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$.*

Proof We present a slight simplification[6] of the original proof. Let P uniformly center its probability mass on d/ϵ arbitrarily chosen instances of X. Let f be the target concept that assigns 0 to all instances. Then L returns a hypothesis of error at most ϵ iff the random sample S contains at least $n - d$ distinct examples. Let Z be the random variable that counts the number of random examples (= trials) which are required in order to get $n - d$ distinct ones. We can decompose the total number of trials into $n - d$ phases where phase i ends immediately after we have seen i distinct examples.

5. Note that the consistency of $h_{j(k)}$ is with $S_1 \cup \ldots \cup S_{j(k)}$ is actually more than we need here: consistency with $S_{j(k)}$ would suffice.
6. We eploit a connection to the so-called Coupon-Collector problem.

Let Z_i be the number of trials in phase i. Then $Z = \sum_{i=1}^{n-d} Z_i$ and Z_i is geometrically distributed with parameter $p_i = (n-i+1)/n$ so that $\mathbb{E}[Z_i] = 1/p_i$ and $\mathrm{Var}[Z_i] = (1-p_i)/p_i^2$. It follows that

$$\mathbb{E}[Z] = n \sum_{i=1}^{n-d} \frac{1}{n-i+1} = n(H_n - H_d)$$

where H_n denotes the n-th Harmonic number. Similarly,

$$\mathrm{Var}[Z] = n \sum_{i=1}^{n-d} \frac{i-1}{(n-i+1)^2} = n \sum_{i=d+1}^{n} \frac{n-i}{i^2} < n^2 \sum_{i=d+1}^{n} \frac{1}{i^2} < \frac{n^2}{d} \ .$$

Let $m = m_L(\epsilon, 1/2)$. It follows that $\Pr[Z \leq m] \geq 1/2$. On the other hand, we infer from the Chebychev inequality that

$$\Pr[Z \leq \mathbb{E}[Z] - \sqrt{2/d}n] \leq \frac{d\mathrm{Var}[Z]}{2n^2} < \frac{1}{2} \ .$$

Making use of $\ln(n) < H_n < 1 + \ln(n)$ and $n = d/\epsilon$, it follows that

$$m > \mathbb{E}[Z] - \sqrt{\frac{2}{d}}n > n(H_n - H_d) - \frac{\sqrt{2d}}{\epsilon} > n\left(\ln\left(\frac{n}{d}\right) - 1\right) - \frac{\sqrt{2d}}{\epsilon} = \frac{d}{\epsilon}\left(\ln\left(\frac{1}{\epsilon}\right) - 1\right) - \frac{\sqrt{2d}}{\epsilon} \ ,$$

which concludes the proof. ■

However, increasing K from 1 to 2 makes L_K optimal for the same class as in the preceding theorem:

Theorem 7 *Let C be the same concept class as in Theorem 6. Let $L : (X \times \{0,1\})^* \to C_d$ be any consistent function. Then the estimation error and the sample complexity function of L_2 are bounded from above as follows:*

$$\epsilon_{L_2}(m, \delta) \leq \frac{18}{m}\left(2d + \ln\left(\frac{9}{\delta}\right)\right) \ \textit{and} \ m_{L_2}(\epsilon, \delta) \leq \frac{18}{\epsilon}\left(2d + \ln\left(\frac{9}{\delta}\right)\right) \ . \tag{15}$$

Proof The proof of Theorem 7 basically proceeds as the proof of Theorem 5. Note however that the error sets E_k (for $k = 1, 2, 3$) can contain at most $2d$ elements because both, the target concept and the hypothesis h_k assign the label 1 to at most d instances of X. As in the proof of Theorem 5, the main challenge is to control the error terms $\epsilon_k = P(E'_k)$ where here k ranges only from 1 to 2. Since the conditional probability measure $P(\cdot|E'_1)$ centers its probability mass on $E'_1 = E_{j(1)}$, a set with at most $2d$ elements, we may apply Theorem 1 instead of Theorem 2 when it comes to bound the estimation error of $h_{j(2)}$ with respect to $P(\cdot|E'_1)$. Note that C_d restricted to E'_1 contains less than 2^{2d} elements so that the term $\ln(|C_d|/\delta)$ within Theorem 1 is upper-bounded by $2d + \ln(1/\delta)$. The "bad events" B_1, B'_2 are defined as in the proof of Theorem 5. However, with an application of Theorem 1 in mind, the event B_2 can now be defined as follows:

$$B_2 \ \Leftrightarrow \ \left(m_2 \geq \frac{1}{2}\epsilon_1 m\right) \wedge \left(\exists g \in V_C(S'_2) : P[g(x) \neq f(x) | x \in E'_1] > \frac{1}{\epsilon_1 m/2}\left(2d + \ln\left(\frac{1}{\delta}\right)\right)\right) \ .$$

According to Theorem 1, the probability of B_2 is bounded by δ. This little modification affects the remainder of the proof only in so far as the error terms ϵ_1, ϵ_2 are now given by $\epsilon_1 \leq \epsilon_{ub}(m, d, \delta)$ and

$$\epsilon_2 \leq \epsilon_1 \cdot \frac{1}{\epsilon_1 m/2}\left(2d + \ln\left(\frac{1}{\delta}\right)\right) = \frac{2}{m}\left(2d + \ln\left(\frac{1}{\delta}\right)\right) \quad .$$

Analogous to our reasoning in the proof of Theorem 5 (but here for the special case $K = 2$ so that $P(E_J) = \epsilon_2$), we may draw the following conclusion: with a probability of at least $1 - \binom{2K-1}{K}(2K-1)\delta = 1 - 9\delta$ (taken over the random selection of the sample), the error probability of the majority vote is bounded by $\binom{2K-1}{K}\frac{2}{m}\left(2d + \ln\left(\frac{1}{\delta}\right)\right) = \frac{6}{m}\left(2d + \ln\left(\frac{1}{\delta}\right)\right)$. In other words,

$$\epsilon_{L_2}(3m, 9\delta) \leq \frac{6}{m}\left(2d + \ln\left(\frac{1}{\delta}\right)\right) \quad .$$

From this, (15) is obtained by a straightforward calculation. ■

5. Final Remarks

We managed to reduce the gap between the general upper bound and the general lower bound on the sample size in the PAC-learning framework (the realizable case) from $\log(1/\epsilon)$ to $\ell_K(1/\epsilon)$ where the constant K can be chosen as large as we like. Evidently, L_K comes very close to Warmuth's concept of an "optimal PAC algorithm" (even for moderate values of K). We now close the paper by discussing some algorithmic modifications of the family $(L_K)_{K\geq 1}$ and some efficiency issues, and by briefly raising some open questions.

Algorithmic modifications. Let L'_K denote the modification of L_K where, for $k = 1, \ldots, 2K-1$, we set $h_k = L(S_k)$ (as opposed to $h_k = L(S_1 \cup \ldots \cup S_k)$ which was the setting of algorithm L_K). Then Theorem 5 and the bounds (13), (14) remain valid when we substitute L'_K for L_K. The reason is that our analysis did not require h_k to be consistent with $S_1 \cup \ldots \cup S_k$. Actually the consistency with S_k is enough. (Compare with the corresponding footnote within the proof of Theorem 5).
Let $\log^*(m)$ be the smallest number K such that $\ell_K(m) < 1$. An alternative to the family $(L_K)_{K\geq 1}$ of algorithms (with one algorithm for each fixed value of K) is obviously the following algorithm L_* which adjusts the choice of K to the size m of the given sample S: set K equal to $\log^*(m)$ and run L_K on S. With this choice of K, the bounds (13) and (14) are still valid.

Efficiency issues. Let $C = (C_d)_{d\geq 1}$ be a parametrized concept class where d is polynomially related to the VC-dimension of C_d. It is well known that C is efficiently properly PAC-learnable iff the consistency problem for C — given d and given a sample S, find a hypothesis $h \in C_d$ that is consistent with S (if possible) — is feasible (Pitt and Valiant, 1988). A nice feature of the algorithms L_K is that they do not need more than the mere feasibility of the consistency problem for running efficiently on C. As for the 1-inclusion graph algorithm, the situation is less comfortable because the size of the 1-inclusion graph grows exponentially with the VC-dimension.

The following questions are left open in this paper:

- Is L_K for $K \geq 2$ even better than it is suggested by our worstcase analysis? Are there (matching) lower bounds?

- Can we combine the hypothesis $(h_k)_{k=1,\ldots,2K-1}$ in a more clever way than just returning the majority vote? Is the experience found in the boosting literature (Schapire and Freund, 2012) helpful for this purpose?
- How does L_K compare experimentally to consistent algorithms?

References

Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2–3):151–163, 2007.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association on Computing Machinery*, 36 (4):929–965, 1989.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

Malte Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2014.

Malte Darnstädt, Balázs Szörényi, and Hans U. Simon. Supervised learning and co-training. *Theoretical Computer Science*, 519:68–87, 2014.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Steve Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0,1\}$ functions on randomly drawn points. *Information and Computation*, 115(2):284–293, 1994.

Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the Association on Computing Machinery*, 35(4):965–984, 1988.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Manfred Warmuth. The optimal PAC algorithm. In *Proceedings of the 17th Annual Conference on Learning Theory*, pages 641–642, 2004.