Smart PAC-Learners *

Hans Ulrich Simon

Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
hans.simon@rub.de

Abstract. The PAC-learning model is distribution-independent in the sense that the learner must reach a learning goal with a limited number of labeled random examples without any prior knowledge of the underlying domain distribution. In order to achieve this, one needs generalization error bounds that are valid uniformly for every domain distribution. These bounds are (almost) tight in the sense that there is a domain distribution which does not admit a generalization error being significantly smaller than the general bound. Note however that this leaves open the possibility to achieve the learning goal faster if the underlying distribution is "simple". Informally speaking, we say a PAC-learner L is "smart" if, for a "vast majority" of domain distributions D, L does not require significantly more examples to reach the "learning goal" than the best learner whose strategy is specialized to D. In this paper, focusing on sample complexity and ignoring computational issues, we show that smart learners do exist. This implies (at least from an information-theoretical perspective) that full prior knowledge of the domain distribution (or access to a huge collection of unlabeled examples) does (for a vast majority of domain distributions) not significantly reduce the number of labeled examples required to achieve the learning goal.

1 Introduction

We are concerned with sample-efficient strategies for properly PAC-learning a finite class $\mathcal{H}$ of concepts over a finite domain X. In the general PAC-learning framework, a learner is exposed to a worst-case analysis by asking the following question: what is the smallest sample size $m = m_{\mathcal{H}}(\varepsilon, \delta)$ such that, for every target concept $h \in \mathcal{H}$ and every distribution D on X, the probability (taken over m randomly chosen and correctly labeled examples) for returning an ε-accurate hypothesis is at least $1 - \delta$? It is well-known [3], [4] that (up to some logarithmic factors) there are matching upper and lower bounds on $m_{\mathcal{H}}(\varepsilon, \delta)$. The proof for the lower bound makes use of a fiendish distribution D^*_ε which makes the learning task quite hard. The lower bound is remains valid when D^*_ε is known to the learner. While this almost completely determines the sample size that is required in the worst-case, it leaves open the question whether the learning goal can be achieved faster when the underlying domain distribution D is significantly simpler than D^*_ε. Furthermore, if it can be achieved faster, it

* This work was supported by the Deutsche Forschungsgemeinschaft Grant SI 498/8-1.

leaves open the question whether this can be exploited only by a learner that is specialized to D or if it can be as well exploited by a "smart" PAC-learner who has no prior knowledge about D. This is precisely the question that we try to answer in this paper.

Our main result: In our paper, it will be convenient to think of the target class $\mathcal{H}$, the sample size m and the accuracy parameter ε as fixed, respectively, and to figure out the smallest value for δ such that m examples suffice to meet the (ε, δ)-criterion of PAC-learning.[1] Let $\delta^*_D = \delta^*_D(\varepsilon, m)$ denote the smallest value for δ that can be achieved by a learner who is specialized to D. A general PAC-learner who must cope with arbitrary distributions can clearly not be specialized to every distribution at the same time. Nevertheless we can associate a quantity $\delta^L_D = \delta^L_D(\varepsilon, m)$ with L that is defined as the smallest value of δ that such that m examples are sufficient to meet the (ε, δ)-criterion provided that L was exposed to the domain distribution D. Ideally, we would like to prove that there is a PAC-learner L and a constant c (not even depending on $\mathcal{H}$) such that, for every domain distribution D, $\delta^L_D \leq c \cdot \delta^*_D$. This would be a strong result as it is basically saying that, for every D, the learning progress of L (without prior knowledge of D) is made roughly at the same speed as the learning progress of the learner whose strategy is perfectly specialized to D. Our result is however slightly weaker than that. We can show that there is a PAC-learner L such that, for a "vast majority" of domain distributions D, $\delta^L_D \leq 2c \cdot \delta^*_D$ (where c is a constant that grows with the desired "vastness" of the majority). The formal statement is found in Corollary 3.

Related papers: Our main result makes a contribution to a discussion about the power of semi-supervised learning that was raised in [1]. The authors pursue the question whether unlabeled data (which are usually cheap) provably help to save labeled data (which are usually expensive). Moreover, they pursue this question in a passive learning model (like PAC-learning) where the labeled data are generated at random (and the learner has no control about the data-generation process). They present some particular concept classes for which they can prove that unlabeled data do not significantly help. Their analysis uses a nice "rescaling trick" that works however only for one-dimensional Euclidean domains. They conjecture that unlabeled data do not significantly help for a much wider family of concept classes. Our main result supports this conjecture for proper learning (as opposed to agnostic learning) and for (arbitrary!) finite classes.
Our results are also weakly related to [2] where upper and lower bounds (in terms of cover- and packing-numbers associated with $\mathcal{H}$ and D) on the sample size are presented when $\mathcal{H}$ is PAC-learned under a fixed distribution D. If these bounds were tight, one line of attack for proving our results could have been to design a general PAC-learner that, when exposed to D, achieves the learning goal with a sample size that does not significantly exceed the lower bound on the sample size in the fixed-distribution setting. However, since the upper and lower

[1] This is obviously equivalent to discussing the sample size as a function in ε and δ.

bounds in [2] leave a significant gap (although being related by a polynomial of small degree), this approach does not look very promising.

Structure of the paper: Section 2 clarifies the notation that is used throughout the paper. Section 3 is devoted to learning under a fixed distribution D. This setting is cast as a zero-sum game between two players, the learner and her opponent, such that the Minimax Theorem from game theory applies. This leads to a nice characterization of $\delta_D^* = \delta_D^*(\varepsilon, m)$. It is furthermore shown that, when the opponent makes his draw first, there is a strategy for the learner that, despite of being not defined in terms of D, does not perform much worse than the best strategy that is specialized to D. Section 4 is devoted to the proof of the main result. To this end, we first treat the case of finitely many distributions and cast the resulting task for a general PAC-learner again as a zero-sum game. Another application of the Minimax Theorem brings us then in the position to prove an important result for a learner who simultaneously copes with finitely many distributions. The case of (infinitely many) arbitrary distributions is finally treated in a similar fashion by invocation of a continuity argument. At the end of Section 4, the reader will find our main results. Section 5 is devoted to some final discussions and open problems.

2 Notations

We assume that the reader is familiar with the PAC-learning framework and knows the Minimax Theorem from game-theory.

Throughout the paper, we use the following notation:

- X denotes a finite domain.
- $\mathcal{H}$ denotes a finite concept class over domain X. Thus, every $h \in \mathcal{H}$ is a function of the form $h : X \to \{0, 1\}$.
- D denotes a domain distribution.
- m denotes the sample size.
- ε denotes the accuracy parameter.
- δ denotes the confidence parameter which bounds from above the "expected failure rate" where "failure" means the delivery of an ε-inaccurate hypothesis.
- $(x, b) \in X^m \times \{0, 1\}^m$ denotes a labeled sample.
- For $x = (\xi_1, \ldots, \xi_m) \in X^m$ and $h \in \mathcal{H}$, we set

$$h(x) = (h(\xi_1), \ldots, h(\xi_m)) \in \{0, 1\}^m \ .$$

- As usual, a learning function is a mapping of the form

$$\mathcal{L} : X^m \times \{0, 1\}^m \to \mathcal{H} \ ,$$

  i.e., it maps labeled samples to hypotheses.

– For two hypotheses $h, h' \in \mathcal{H}$,

$$h \oplus h' := \{\xi \in X : \ h(\xi) \neq h'(\xi)\}$$

denotes their symmetric difference. Recall that h is called ε-accurate for h' w.r.t. D if $D(h \oplus h') \leq \varepsilon$.

A deterministic learner can be identified with a learning function (if computational issues are ignored). We consider, however, randomized learners. Eachone of these can be identified with a probability distribution over the set of all learning functions.

3 Learning Under a Fixed Distribution Revisited

Consider a fixed distribution D on X that is known to the learner. Let $\mathcal{L}_1, \ldots, \mathcal{L}_M$ be a list of all learning functions mapping a labeled sample of size m to a hypothesis from $\mathcal{H} = \{h_1, \ldots, h_N\}$. For every $\varepsilon > 0$, $i = 1, \ldots, M$, $j = 1, \ldots, N$, $x \in X^m$, and $b \in \{0,1\}^m$, let $I_D^{\varepsilon,x,b}[i,j]$ be the Bernoulli variable indicating whether the hypothesis $\mathcal{L}_i(x,b)$ is ε-inaccurate w.r.t. target concept h_j and domain distribution D, i.e.,

$$I_D^{\varepsilon,x,b}[i,j] = \begin{cases} 1 \text{ if } \mathcal{L}_i(x,b) \text{ is } \varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D \\ 0 \text{ otherwise} \end{cases} .$$

Now, let

$$A_D^{\varepsilon,m}[i,j] := \mathbb{E}_{x \in D^m}\left[I_D^{\varepsilon,x,h_j(x)}[i,j]\right] = \sum_x D^m(x) I_D^{\varepsilon,x,h_j(x)}[i,j] \tag{1}$$

$$= \Pr_{x \in D^m}\left[\mathcal{L}_i(x, h_j(x)) \text{ is } \varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D\right] . \tag{2}$$

If D, ε, m are obvious from context, we omit these letters as subscripts or superscripts in what follows. A randomized learner is given by a vector $p \in [0,1]^M$ that assigns a probability p_i to every learning function $\mathcal{L}_i$ (so that $\sum_{i=1}^M p_i = 1$). Thus, we may identify learners with mixed strategies for the "row-player" in the zero-sum game associated with A. We may view the "column-player" in this game as an opponent of the learner. A mixed strategy for the opponent is given by a vector $q \in [0,1]^N$ that assigns an à-priori probability q_j to every possible target concept h_j (so that $\sum_{j=1}^N q_j = 1$). The well-known Minimax Theorem states that

$$\min_p \max_q p^\top A q = \max_q \min_p p^\top A q \ . \tag{3}$$

In the sequel, we denote the optimal value in (3) by $\delta_D^*(\varepsilon, m)$. Note that $\delta_D^*(\varepsilon, m)$ coincides with the smallest value for the confidence parameter δ such that every target concept from $\mathcal{H}$ can be inferred up to accuracy ε from a random sample of size m (w.r.t. the fixed distribution D) with a probability at least $1 - \delta$ of success. Thus the definition of $\delta_D^*(\varepsilon, m)$ given here is consistent with the definition given in the introduction.

Recall that by "expected failure rate (w.r.t. D, m, ε)" we mean the probability for delivering an ε-inaccurate hypothesis. We denote by $\delta_D^p(\varepsilon, m)$ the expected failure rate of the learner with mixed strategy p (with the opponent making the second draw). Clearly,

$$\delta_D^*(\varepsilon, m) = \min_p \delta_D^p(\varepsilon, m) = \min_p \max_{j=1,\ldots,N} p^\top A_j$$

where A_j denotes the j'th column of matrix $A = A_D^{\varepsilon,m}$.

Since a mixed strategy for the learner is a distribution over learning functions (mapping a labeled sample to a hypothesis), we may equivalently think of the learner as waiting for a random labeled sample (x, b) and then playing a mixed strategy that depends on (x, b). In order to formalize this intuition, we consider the new payoff-matrix $\tilde{A} = \tilde{A}_D^\varepsilon$ given by

$$\tilde{A}[i, j] = \begin{cases} 1 \text{ if } h_i \text{ is } \varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D \\ 0 \text{ otherwise} \end{cases} .$$

We associate the following game with $\tilde{A}$:

1. The opponent selects a vector $q \in [0, 1]^N$ specifying à-priori probabilities for the target concept. Note that this implicitly determines
   - the probability
   $$Q(b|x) = \sum_{j:h_j(x)=b} q_j$$
   of labeling a given sample x by b,
   - and the à-posteriori probabilities
   $$Q(j|x, b) = \begin{cases} \frac{q_j}{Q(b|x)} \text{ if } h_j(x) = b \\ 0 \quad \text{otherwise} \end{cases} \tag{4}$$
   for target concept h_j given the labeled sample (x, b).

   For sake of a compact notation, let $\tilde{q}(x, b)$ denote the vector whose j'th component is $Q(j|x, b)$.
2. A labeled sample (x, b) is produced at random with probability
   $$\Pr(x, b) = D^m(x)Q(b|x) \ . \tag{5}$$
3. The learner chooses a vector $\tilde{p}(x, b) \in [0, 1]^N$ (that may depend on D, q and (x, b)) specifying her mixed strategy w.r.t. payoff-matrix $\tilde{A}$.
4. The learner suffers loss $\tilde{p}(x, b)^\top \tilde{A}\tilde{q}(x, b)$ so that her expected loss, averaged over all labeled samples, evaluates to $\sum_{x,b} \Pr(x, b)\tilde{p}(x, b)^\top \tilde{A}\tilde{q}(x, b)$.

In the sequel, the games associated with A and $\tilde{A}$, respectively, are simply called A-game and $\tilde{A}$-game, respectively.

Lemma 1. *Let $q \in [0,1]^N$ be an arbitrary but fixed mixed strategy for the learner's opponent. Then every mixed strategy $p \in [0,1]^M$ for the learner in the A-game can be mapped to a mixed strategy $\tilde{p}$ for the learner in the $\tilde{A}$-game so that*

$$p^\top A q = \sum_{x,b} \Pr(x,b) \tilde{p}(x,b)^\top \tilde{A} \tilde{q}(x,b) \ . \tag{6}$$

Moreover, this mapping $p \mapsto \tilde{p}$ is surjective, i.e., every mixed strategy for the learner in the $\tilde{A}$-game has a pre-image (so that the optimal values in both games are the same).

Proof. For every probability vector $p \in [0,1]^M$ and every labeled sample (x,b), we define the corresponding probability vector $\tilde{p}(x,b) \in [0,1]^N$ as follows:

$$\tilde{p}_{i'}(x,b) = \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} p_i \tag{7}$$

Note that

$$\sum_{i'=1}^{N} \tilde{p}_{i'}(x,b) = \sum_{i=1}^{M} p_i = 1 \ .$$

The following computation verifies (6):

$$\begin{aligned}
p^\top A q &\overset{(1)}{=} \sum_x D^m(x) \sum_{j=1}^{N} \sum_{i=1}^{M} I^{x,h_j(x)}[i,j] p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i=1}^{M} I^{x,b}[i,j] p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i'=1}^{N} \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} \underbrace{I^{x,b}[i,j]}_{=\tilde{A}[i',j]} p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i'=1}^{N} \tilde{A}[i',j] q_j \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} p_i \\
&\overset{(7)}{=} \sum_{x,b} D^m(x) \sum_{i'=1}^{N} \sum_{j:h_j(x)=b} \tilde{A}[i',j] \tilde{p}_{i'}(x,b) q_j \\
&\overset{(4)}{=} \sum_{x,b} D^m(x) Q(b|x) \sum_{i'=1}^{N} \sum_{j=1}^{N} \tilde{A}[i',j] \tilde{p}_{i'}(x,b) Q(j|x,b) \\
&\overset{(5)}{=} \sum_{x,b} \Pr(x,b) \tilde{p}(x,b)^\top \tilde{A} \tilde{q}(x,b)
\end{aligned}$$

As for the second part of the lemma, consider a mixed strategy $\tilde{p}'(x,b)$ of the learner in the $\tilde{A}$-game. We shall specify a mixed strategy p of the learner in

the A-game such that function $\tilde{p}(x,b)$ computed according to (7) coincides with $\tilde{p}'(x,b)$. To this end, let us make the notational convention $\tilde{p}'_{h_{i'}} := \tilde{p}'_{i'}$ and let us choose p as follows:

$$p_i = \prod_{x,b} \tilde{p}'_{\mathcal{L}_i(x,b)}(x,b) \ . \tag{8}$$

Note that, with this choice of p,

$$\sum_{i=1}^{M} p_i \stackrel{(8)}{=} \sum_{i=1}^{M} \prod_{x,b} \tilde{p}'_{\mathcal{L}_i(x,b)}(x,b) = \prod_{x,b} \underbrace{\sum_{i'=1}^{N} \tilde{p}'_{i'}(x,b)}_{=1} = 1 \ . \tag{9}$$

In the second-last equation, we used the distributive law where the reader should note the one-to-one correspondence between the set of all learning functions and the free combination of (number of labeled samples many) hypotheses taken from $\mathcal{H}$.
A computation similar to (9) verifies now the desired coincidence between $\tilde{p}$ and $\tilde{p}'$:

$$\begin{aligned}
\tilde{p}_{i'}(x',b') &\stackrel{(7)}{=} \sum_{i:\mathcal{L}_i(x',b')=h_{i'}} p_i \\
&\stackrel{(8)}{=} \sum_{i:\mathcal{L}_i(x',b')=h_{i'}} \prod_{x,b} \tilde{p}'_{\mathcal{L}_i(x,b)}(x,b) \\
&= \tilde{p}'_{i'}(x',b') \sum_{i:\mathcal{L}_i(x',b')=h_{i'}} \prod_{(x,b)\neq(x',b')} \tilde{p}'_{\mathcal{L}_i(x,b)}(x,b) \\
&= \tilde{p}'_{i'}(x',b') \prod_{(x,b)\neq(x',b')} \underbrace{\sum_{i=1}^{N} \tilde{p}'_{i}(x,b)}_{=1} \\
&= \tilde{p}'_{i'}(x',b')
\end{aligned}$$

In the second-last equation, we used the distributive law where the reader should note the one-to-one correspondence between the set of all learning functions with a fixed value on one sample (x',b') and the free combination of (number of labeled samples minus 1 many) hypotheses taken from from $\mathcal{H}$. □

As a corollary to Lemma 1, we obtain

$$\delta^*_D(\varepsilon,m) = \max_q \min_p \, p^\top A_D^{\varepsilon,m} q = \max_q \left[\sum_{x,b} \left(\Pr(x,b) \cdot \min_{\tilde{p}} \left[\tilde{p}^\top \tilde{A}_D^{\varepsilon,m} \tilde{q}(x,b) \right] \right) \right] \ .$$

We close this section with a result that prepares the ground for our analysis of general PAC-learners in the next section:

Lemma 2. *Let $\varepsilon > 0$ be a given accuracy and $m \geq 1$ a given sample size. For every probability vector $q \in [0,1]^N$, and every domain distribution D, the following holds:*

$$\sum_{x,b} \Pr(x,b)\tilde{q}(x,b)^\top \tilde{A}_D^{2\varepsilon}\tilde{q}(x,b) \leq 2\delta_D^*(\varepsilon,m) \tag{10}$$

Proof. Recall that $\tilde{q}(x,b)$ is the vector that assigns the à-posteriori probability $Q(j|(x,b)$ for being the target concept to every hypothesis h_j. Since the à-posteriori probabilities outside the version space

$$V := \{h \in \mathcal{H} :\ h(x) = b\}$$

are zero, only target concepts in V can contribute to the left hand-side in (10). In the remainder of the proof, we simply write $\tilde{A}^\varepsilon$ instead of $\tilde{A}_D^\varepsilon$, and $\tilde{A}_i^\varepsilon$ denotes the i'th row of this matrix. In the $\tilde{A}^\varepsilon$-game, the opponent makes the first draw by choosing a (prior) probability vector $q \in [0,1]^N$. The following "Bayesian strategy" for the learner minimizes (6): for a given labeled sample (x,b) pick a hypothesis $h^* = h_{i^*(x,b)} \in \mathcal{H}$ which maximizes the total à-posteriori probability of hypotheses that are ε-close to h^* w.r.t. D, i.e.,

$$i^*(x,b) = \arg\max_{i=1,\ldots,N} \left\{ \sum_{j:D(h_i \oplus h_j) \leq \varepsilon} Q(j|x,b) \right\} \ .$$

It follows that

$$\sum_{x,b} \Pr(x,b)\tilde{A}^\varepsilon_{i^*(x,b)}\tilde{q}(x,b) \leq \delta_D^*(\varepsilon,m) \ . \tag{11}$$

We are now prepared to verify (10). We call a hypothesis from the version space V an "(D,x,b,ε)-exception" if it is not ε-close to h^* w.r.t. D. Note that $\tilde{A}^\varepsilon_{i^*(x,b)}\tilde{q}(x,b)$ coincides with the total à-posteriori probability of (D,x,b,ε)-exceptions. Consider now the strategy $\tilde{p} = \tilde{q}$ for the learner. Given (x,b), she picks a hypothesis $\hat{h} = h_i$ at random with probability $Q(i|x,b)$. The following observation, which is a simple application of the triangle inequality, is crucial: if $\hat{h} \in V$ is not an (D,x,b,ε)-exception, then $\hat{h}$ is 2ε-close w.r.t. D to every hypothesis in V that is not an (D,x,b,ε)-exception either. Thus, if $\hat{h}$ and the target concept are both picked at random according to $\tilde{q}(x,b)$, then the probability for $\hat{h}$ being 2ε-inaccurate is bounded from above by twice the total probability of (D,x,b,ε)-exceptions, i.e.,

$$\tilde{q}(x,b)\tilde{A}^{2\varepsilon}\tilde{q}(x,b) \leq 2A^\varepsilon_{i^*(x,b)}\tilde{q}(x,b) \ .$$

This, combined with (11), concludes the proof of the lemma. □

It is important to note that no knowledge of D is required to play the strategy $\tilde{p} = \tilde{q}$ in the $\tilde{A}$-game (with the opponent making the first draw). Nevertheless, as made precise in Lemma 2, this is a reasonably good strategy for *any* underlying domain distribution.

4 Smart PAC-Learners

Let us first consider learners that cope with an arbitrary but fixed finite list $D_1, \ldots, D_C$ of (not necessarily distinct) distributions on X.[2] We shall define a suitable payoff-matrix R blockwise so that $R = [R^{(1)}, \ldots, R^{(C)}]$ and $R^{(k)}$ is the block reserved for distribution D_k. Every block has M rows (one row for every learning function) and N columns (one column for every possible target concept). Choosing $R^{(k)} = A^{\varepsilon,m}_{D_k}$ (compare with the previous section) would lead to mixed strategies for the PAC-learner that put too much emphasis on fiendish domain distributions. Such strategies are not likely to succeed with considerably fewer sample points when the underlying distribution D happens to be simple. Assuming

(A1) $\forall k = 1, \ldots, C,\ \delta^*_{D_k}(\varepsilon, m) \neq 0,$

the following payoff-matrix is a better choice:

$$\begin{aligned} R^{(k)}[i,j] &:= \frac{1}{\delta^*_{D_k}(\varepsilon,m)} \cdot A^{2\varepsilon,m}_{D_k}[i,j] \\ &= \frac{1}{\delta^*_{D_k}(\varepsilon,m)} \Pr_{x \in D_k^m}[\mathcal{L}_i(x, h_j(x)) \cdot \text{ is } 2\varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D_k] \end{aligned}$$

Intuitively, the scaling factor $1/\delta^*_{D_k}(\varepsilon, m)$ challenges the learner to put more emphasis on benign distributions. Note furthermore that we penalize the learner only if her hypothesis is 2ε-inaccurate (as opposed to ε-inaccurate). This leaves some slack which helps the learner to compensate for not knowing D.

In the sequel, we simply write $\delta^*_k(\varepsilon, m)$ instead of $\delta^*_{D_k}(\varepsilon, m)$ and $A^{(k)}$ instead of $A^{2\varepsilon,m}_{D_k}$. $A^{(k)}_j$ denotes the j'th column in $A^{(k)}$ and, similarly, $R^{(k)}_j$ denotes the j'th column in $R^{(k)}$. A mixed strategy for the learner is a probability vector $p \in [0,1]^M$ (as in Section 3). A mixed strategy for the opponent is a probability vector $\boldsymbol{q} = [q^{(1)}, \ldots, q^{(C)}] \in [0,1]^{CN}$ where $q^{(k)}_j$ denotes the probability for choosing domain distribution D_k and target concept h_j. According to the Minimax Theorem, the following holds:

$$\min_p \max_{\boldsymbol{q}} p^\top R \boldsymbol{q} = \max_{\boldsymbol{q}} \min_p p^\top R \boldsymbol{q} \ . \tag{12}$$

Let $\rho^*(\varepsilon, m)$ denote the optimal value in (12). The following quantities refer to a learner who makes the first draw and applies the mixed strategy p:

$$\rho^p(\varepsilon, m) := \max_{j=1,\ldots,N} \max_{k=1,\ldots,C} p^\top R^{(k)}_j$$

$$\bar{\rho}^p(\varepsilon, m) := \max_{j=1,\ldots,N} \frac{1}{C} \cdot \sum_{k=1}^{C} p^\top R^{(k)}_j$$

Clearly, $\rho^*(\varepsilon, m) = \min_p \rho^p(\varepsilon, m)$.

[2] We shall later extend these considerations to arbitrary distributions.

Let us make perfectly clear the connection between these quantities and our concept of a smart PAC-learner. We denote by $\delta^p_{j,k}(\varepsilon, m)$ the expected failure rate (w.r.t. m, ε) of a PAC-learner with mixed strategy p when the target concept is h_j and the domain distribution is D_k.[3] It follows from the definition of $A^{(k)} = A^{2\varepsilon,m}_{D_k}$ that

$$\delta^p_{j,k}(2\varepsilon, m) = p^\top A^{(k)}_j \quad .$$

Thus, according to the definition of $R^{(k)}$,

$$\begin{aligned}
\frac{\delta^p_{j,k}(2\varepsilon, m)}{\delta^*_k(\varepsilon, m)} &= p^\top R^{(k)}_j \quad , \\
\max_{j=1,\ldots,N} \max_{k=1,\ldots,C} \frac{\delta^p_{j,k}(2\varepsilon, m)}{\delta^*_k(\varepsilon, m)} &= \rho^p(\varepsilon, m) \quad , \\
\max_{j=1,\ldots,N} \frac{1}{C} \cdot \sum_{k=1}^{C} \frac{\delta^p_{j,k}(2\varepsilon, m)}{\delta^*_k(\varepsilon, m)} &= \bar{\rho}^p(\varepsilon, m) \quad .
\end{aligned}$$

It becomes obvious now that the quantities $\rho^p(\varepsilon, m)$ and $\bar{\rho}^p(\varepsilon, m)$ measure how well the general PAC-learner with mixed strategy p (and accuracy parameter 2ε) competes against the best learner with full prior knowledge of the domain distribution (and accuracy parameter ε). We call $\rho^p(\varepsilon, m)$ the *worst performance ratio* and $\bar{\rho}^p(\varepsilon, m)$ the *average performance ratio* of the mixed strategy p (although both quantities refer to the worst-case as far as the choice of the target concept is concerned). In the sequel, a learner with mixed strategy p is identified with p so that we can speak of a performance ratio of the learner. A very strong result would be $\rho^*(\varepsilon, m) \leq c$ for some small constant c, which would mean that there exists a learner (mixed strategy) p whose worst performance ratio is bounded by c. But, since this is somewhat overambitious, we pursue a weaker goal in the following and analyze the average performance ratio, $\bar{\rho}^p(\varepsilon, m)$, instead. We make use of the (obvious) fact that

$$\bar{\rho}^*(\varepsilon, m) := \min_p \bar{\rho}^p(\varepsilon, m)$$

is the optimal value in the $\bar{R}$-game for

$$\bar{R} := \frac{1}{C} \sum_{k=1}^{C} R^{(k)} \quad . \tag{13}$$

With this notation, the following holds:

Lemma 3. *For every mixed strategy q of the opponent in the $\bar{R}$-game, there exists a mixed strategy p for the learner such that $p^\top \bar{R} q \leq 2$.*

[3] In contrast to the previous section, here the learner has no prior knowledge of D_k.

Proof. From the decomposition (6), we get the following decomposition of $p^\top \bar{R} q$:

$$\begin{aligned} p^\top \bar{R} q &\overset{(13)}{=} \frac{1}{C} \sum_{k=1}^{C} p^\top R^{(k)} q \\ &= \frac{1}{C} \sum_{k=1}^{C} \frac{1}{\delta^*_{D_k}(\varepsilon, m)} p^\top A^{(k)} q \\ &\overset{(6)}{=} \frac{1}{C} \sum_{k=1}^{C} \frac{1}{\delta^*_{D_k}(\varepsilon, m)} \sum_{x,b} \Pr_k(x,b) \tilde{p}(x,b) \tilde{A}^{(k)} \tilde{q}(x,b) \ . \end{aligned}$$

Here, $\Pr_k(x,b) = D_k^m(x) Q(b|x)$, $\tilde{A}^{(k)} = \tilde{A}^{2\varepsilon}_{D_k}$, and the quantities $Q(b|x), \tilde{q}, \tilde{p}$ are derived from q and p, respectively, as explained in the previous section. According to Lemma 2,

$$\sum_{x,b} \Pr_k(x,b) \tilde{q}(x,b) \tilde{A}^{(k)} \tilde{q}(x,b) \leq 2\delta^*_{D_k}(\varepsilon, m) \ . \tag{14}$$

According to Lemma 1, there exists a mixed strategy p for the learner such that $\tilde{p} = \tilde{q}$. With this choice of p, we get

$$p^\top \bar{R} q = \frac{1}{C} \sum_{k=1}^{C} \frac{1}{\delta^*_{D_k}(\varepsilon, m)} \sum_{x,b} \Pr_k(x,b) \tilde{q}(x,b) \tilde{A}^{(k)} \tilde{q}(x,b) \overset{(14)}{\leq} 2 \ ,$$

as desired. □

Let $\bar{R}_j$ denote the j'th column of matrix $\bar{R}$. The Minimax Theorem applied to the $\bar{R}$-game allows us to infer from Lemma 3 the following

Corollary 1. *$\bar{\rho}^*(\varepsilon, m) \leq 2$, i.e., there exists a learner (mixed strategy) p whose average performance ratio is bounded by 2.*

So far, we have assumed that there is a finite list of distributions and the domain distribution is taken from this list. We now extend these considerations to arbitrary distributions. Recall that our domain X is finite, say $X = \{\xi_1, \ldots, \xi_d\}$. The domain distributions are in one-to-one correspondence with the vectors taken from the probability simplex

$$\Delta := \{z \in [0,1]^d : \ z_1 + \cdots + z_d = 1\} \ .$$

Specifically, $D_z(\xi_\nu) = z_\nu$ for $\nu = 1, \ldots, d$. Note that instead of finitely many block matrices $R^{(1)}, \ldots, R^{(C)}$, as before, we now have a system of infinitely many matrices $R^{(z)}$. Let $f : \Delta \to \mathbb{R}_+$ be a continuous function that satisfies

$$\int_{z \in \Delta} f(z) \, dz = 1 \tag{15}$$

so that it can serve as a density function. For sake of simple notation, we set

$$\delta^*_z := \delta^*_{D_z}(\varepsilon, m) \ . \tag{16}$$

For every $\zeta > 0$ and every $E \subseteq \Delta$, let

$$\Pr(E) := \int_{z \in E} f(z)\, dz \ , \tag{17}$$

$$\Delta_\zeta := \{z \in \Delta : \delta_z^* \geq \zeta\} \ . \tag{18}$$

The former Assumption (A1) is replaced now by the following assumption:

(A2) $\lim_{\zeta \to 0} \Pr(\Delta_\zeta) = 1$.

(A2) implies that $\Pr(\Delta_\zeta) > 0$ for every sufficiently small $\zeta > 0$, which is assumed in the sequel. Since $f(z)/\Pr(\Delta_\zeta)$ is a continuous density function on Δ_ζ, we can now use

$$\bar{R}^\zeta := \frac{1}{\Pr(\Delta_\zeta)} \cdot \int_{z \in \Delta_\zeta} f(z) R^{(z)}\, dz \tag{19}$$

as a payoff-matrix (where this matrix-equation is understood entry-wise).

Lemma 4. *The integral on the right hand-side in (19) exists.*

Proof. Since $f(z)$ is continuous and Δ is compact, f is continuous and bounded on Δ. Furthermore, for every $z \in \Delta_\zeta$, each entry of $R^{(z)}$ is bounded by $1/\zeta$. The lemma would be rather obvious if the functions $R^{(z)}[i,j]$ were continuous in z. Although this is not the case in general, we may exploit the fact that discontinuities occur only in the set

$$E = \bigcup_{1 \leq i < j \leq N} E_{ij}$$

where

$$E_{ij} := \{z \in \Delta : \ D_z(h_i \oplus h_j) = \varepsilon\} \ .$$

Note that the sets E_{ij} are of Lebesgue-measure zero, and so is E. For this reason, integrating over Δ_ζ leads to the same result as integrating over $\Delta_\zeta \setminus E$, which, by construction, is a set without discontinuities. □

The *average performance ratio* of a mixed strategy p for the learner refers now to the density function $f(z)/\Pr(\Delta_\zeta)$ and must therefore be redefined as follows:

$$\bar{\rho}_\zeta^p(\varepsilon, m) := \max_{j=1,\ldots,N} \frac{1}{\Pr(\Delta_\zeta)} \cdot \int_{z \in \Delta_\zeta} f(z) p^\top R_j^{(z)}\, dz = \max_{j=1,\ldots,N} p^\top \bar{R}_j^\zeta$$

where $\bar{R}_j^\zeta$ denotes the j'th column of $\bar{R}^\zeta$. With this notation, we get

Corollary 2. *For every sufficiently small $\zeta > 0$, there exists a learner (mixed strategy) p whose average performance ratio is bounded by 2.*

Proof. The crucial observation is that Lemma 3 is still correct when we define $\bar{R} := \bar{R}^\zeta$ according to (19). The only modification in the proof of this lemma is the substitution of integrals for sums. Thus the Minimax Theorem applies to the $\bar{R}$-game and Corollary 2 is obtained. □

Assumption (A2) and Corollary 2 combined with Markov's Inequality immediately lead to

Corollary 3. *For every continuous function $f : \Delta \to \mathbb{R}_+$ satisfying (15) and for every constant $c > 0$, there exists a mixed strategy p for the learner such that, for $j = 1, \ldots, N$,*

$$\Pr(z \in \Delta : p^\top R_j^{(z)} \leq 2c) > 1 - \frac{1}{c} \ . \tag{20}$$

Proof. Let $\zeta > 0$ and $\bar{R} = \bar{R}_\zeta$. Corollary 2 combined with Markov's Inequality shows that there exists a mixed strategy p for the learner such that, for $j = 1, \ldots, N$,

$$\Pr(z \in \Delta_\zeta : p^\top R_j^{(z)} > 2c) < \frac{1}{c} \ .$$

By assumption (A2), we may conclude that

$$\Pr(z \in \Delta : p^\top R_j^{(z)} > 2c) \leq \Pr(z \in \Delta_\zeta : p^\top R_j^{(z)} > 2c) + \Pr(\Delta \setminus \Delta_\zeta) < \frac{1}{c}$$

provided that $\zeta > 0$ is sufficiently small. From this, Corollary 3 is immediate. □

According to Corollary 3, there exists a mixed strategy p for a learner without any prior knowledge of the domain distribution such that, in comparison to the best learner with full prior knowledge of the domain distribution, a performance ratio of $2c$ is achieved for the "vast majority" of distributions. The total probability mass of distributions (measured according to density function $f(z)$) not belonging to the "vast majority" is bounded by $1/c$. So Corollary 3 is the result that we had announced in the introduction.

5 Discussion of Assumption (A2) and Open Problems

We claim that Assumption (A2) is not very restrictive and provide some intuition why this might be true. For every $\gamma > 0$, let

$$\Delta'_\gamma := \{z \in \Delta |\ \forall \nu = 1, \ldots, d : \gamma \leq z_\nu \leq 1 - \gamma\} \ .$$

Assume that $z \in \Delta'_\gamma$. Pick $\nu(z) \in \{1, \ldots, d\}$ such that $z_{\nu(z)} = \min\{z_1, \ldots, z_d\}$. Clearly,

$$\gamma \leq z_{\nu(z)} \leq \frac{1}{d} \ .$$

For sake of brevity, let $\xi := \xi_{\nu(z)}$. For $b \in \{0, 1\}$, consider the set

$$\mathcal{H}(\xi, b) := \{h \in \mathcal{H} : h(\xi) = b\} \ .$$

If the opponent finds two hypotheses $h, h' \in \mathcal{H}(\xi, b)$ such that

$$D_z(h \oplus h') > 2\varepsilon \ , \tag{21}$$

he can assign à-priori probability $1/2$ to h and h', respectively, and achieve the following:
With a probability of at least γ^m, the sample x is of the form $x = (\xi, \ldots, \xi) \in X^m$ so that the learner cannot distinguish between h and h'. Conditioned to $x = (\xi, \ldots, \xi) \in X^m$, the learner will therefore fail with probability at least $1/2$ (regardless of her strategy). Thus, the overall expected failure rate is at least $\gamma^m/2$.
The punchline of this discussion is the following implication:

$$\big(z \in \Delta'_\gamma\big) \wedge \underbrace{\left(2\varepsilon < \max_{b\in\{0,1\}} \max_{g,g' \in \mathcal{H}(\xi_{\nu(z)},b)} D_z(g \oplus g')\right)}_{(A3)} \Rightarrow \delta^*_z(\varepsilon, m) \geq \frac{\gamma^m}{2} \qquad (22)$$

Condition (A3) looks wild but it is essentially saying that the knowledge of a single labeled example should not trivialize the resulting version space (in terms of its diameter) too much.

Define $K(\mathcal{H})$ as the smallest number K such that, for every $\xi \in X$, there exist $g_+ \in \mathcal{H}(\xi, 1)$, $g_- \in \mathcal{H}(\xi, 0)$ and $g_1, \ldots, g_K \in \mathcal{H}$ which satisfy the following condition:

$$\begin{aligned}\forall \xi' \in X \setminus \{\xi\}, \exists \kappa \in \{1, \ldots, K\} : \; & (g_+(\xi) = g_\kappa(\xi) \wedge g_+(\xi') \neq g_\kappa(\xi')) \\ & \vee (g_-(\xi) = g_\kappa(\xi) \wedge g_-(\xi') \neq g_\kappa(\xi'))\end{aligned}$$

If suitable functions $g_+, g_-, g_1, g_2, \ldots$ cannot be found for some $\xi \in X$, we set $K(\mathcal{H}) = \infty$ by default.

Lemma 5. *Assume that $K(\mathcal{H}) < \infty$. Then, for every $z \in \Delta'_\gamma$ and every $d \geq 2$, the following holds:*

$$\max_{b\in\{0,1\}} \max_{g,g' \in \mathcal{H}(\xi_{\nu(z)},b)} D_z(g \oplus g') \geq \frac{1 - 1/d}{K(\mathcal{H})} \geq \frac{1}{2K(\mathcal{H})}$$

Proof. According to the definition of $\nu(z)$, $D_z(X \setminus \{\xi_{\nu(z)}\}) \geq 1 - 1/d$. Let us set $\xi := \xi_{\nu(z)}$, let $K = K(\mathcal{H})$, and let $g_+, g_-, g_1, \ldots, g_K$ be the functions chosen in accordance with the definition of $K(\mathcal{H})$. Without loss of generality, we may assume that $g_1, \ldots, g_{K'}$ agree with g_+ on ξ whereas $g_{K'+1}, \ldots, g_K$ agree with g_- on ξ. Note that the "disagreement sets"

$$g_+ \oplus g_1, \ldots, g_+ \oplus g_{K'} \; ; \; g_- \oplus g_{K'+1}, \ldots, g_- \oplus g_K$$

cover $X \setminus \{\xi\}$. Thus, by the pigeon-hole principle, there must exist a hypothesis $g \in \{g_+, g_-\}$ and $g' \in \{g_1, \ldots, g_K\}$ such that $g(\xi) = g'(\xi)$ but the disagreement set $g \oplus g'$ has probability mass at least $(1 - 1/d)/K$. □

Lemma 5 combined with the implication (22) yields

Corollary 4. *If $K(\mathcal{H}) < \infty$, then the following holds:*

$$\forall \gamma > 0, \forall z \in \Delta'_\gamma, \forall \varepsilon < \frac{1}{2K(\mathcal{H})} : \delta^*_z(\varepsilon, m) \geq \frac{\gamma^m}{2}$$

We remind the reader to (16) and to the definition of Δ_ζ in (18). Since $\gamma^m/2 \geq \zeta$ is equivalent to $\gamma \geq (2\zeta)^{1/m}$, we obtain

Corollary 5. *Let $K(\mathcal{H}) < \infty$ and $\varepsilon < 1/(2K(\mathcal{H}))$. Define $\gamma(\zeta) := (2\zeta)^{1/m}$. Then,*

$$\Delta_\zeta \supseteq \Delta'_{\gamma(\zeta)} \ .$$

Since $\gamma(\zeta) \to 0$ for $\zeta \to 0$ and obviously $\lim_{\gamma \to 0} \Pr(\Delta'_\gamma) = 1$ for every underlying continuous density function, we finally get

Corollary 6. *Let $K(\mathcal{H}) < \infty$. Then, for every $\varepsilon < 1/(2K(\mathcal{H}))$, $\mathcal{H}$ satisfies Assumption (A2).*

A simple hypothesis class should intuitively have a good chance to trivialize the version space resulting from a single labeled example (and, thus, a good chance to violate Assumption (A2)). But even for the almost trivial class of half-intervals $\{1, \ldots, r\}$, $0 \leq r \leq d$, our sufficient condition $K(\mathcal{H}) < \infty$ applies. This can be seen as follows:
Pick an arbitrary $\xi \in \{1, \ldots, d\}$, and designate the following half-intervals:

$$g_+ := \{1, \ldots, \xi\} \ , \ g_- := \emptyset \ ,$$
$$g_1 := \{1, \ldots, d\} \ , \ g_2 := \{1, \ldots, \xi - 1\} \ .$$

Pick an arbitrary $\xi' \in \{1, \ldots, d\} \setminus \{\xi\}$. For $\xi' > \xi$, we get $g_+(\xi) = g_1(\xi) = 1$ and $g_+(\xi') = 0 \neq 1 = g_1(\xi')$. For $\xi' < \xi$, we get $g_-(\xi) = g_2(\xi) = 0$ and $g_-(\xi') = 0 \neq 1 = g_2(\xi')$. We conclude that $K(\mathcal{H}) = 2$ for the class of half-intervals.

Open problems:

- For every finite hypothesis class, we have shown the mere existence of a learner (mixed strategy) whose average performance ratio is bounded by 2. Gain more insight how this strategy actually works and check under which conditions it can be implemented efficiently.
- Prove or disprove that there exists a learner (mixed strategy) whose *worst* performance ratio is bounded by a small constant.
- Prove or disprove our claim that assumption (A2) is not very restrictive.

References

1. Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 33–44, 2008.
2. Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.
3. 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.
4. 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.