Smart PAC-Learners [⋆]

Malte Darnstädt and Hans Ulrich Simon

Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany

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 from $\mathcal{H}$ is at least $1 - \delta$? It is well-known [8], [13]

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

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 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 leaves open the question whether this can be exploited only by a learner who 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 $\delta_D^*(\varepsilon, m)$ 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 $\delta_D^L(\varepsilon, m)$ 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 such that, for every domain distribution D, $\delta_D^L(\varepsilon, m) = \delta_D^*(\varepsilon, m)$. 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 at the same speed as the learning progress of the learner whose strategy is perfectly specialized to D. We will, however, provide some compensation for having no prior knowledge of D and arrive at a slightly weaker result:

Limitation 1: We aim at an inequality of the form

$$\delta_D^L(2\varepsilon, m) \leq c \cdot \delta_D^*(\varepsilon, m) \tag{1}$$

where c denotes a constant (not even depending on $\mathcal{H}$). In other words, L may be twice as inaccurate as the learner who is specialized to D, and its probability to fail may be larger by a constant factor.

Limitation 2: The inequality (1) is not verified for all distributions D but for a volume of approximately $1-2/c$, say when measured with the uniform distribution over the $|X|$-dimensional probability simplex (whose points represent distributions over X). Actually the final result is somewhat more general: we can fix any measure μ on the $|X|$-dimensional probability simplex (expressing which domain distributions we consider as particularly important), and still achieve validity of (1) for a volume of approximately $1-2/c$ of

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

all domain distributions. Clearly, the design of the smart learner L changes when μ changes.

The formal statement is found in Theorem 10.

Relation of our Work to Semi-supervised Learning: Our work is related to the question to which extent unlabeled data (which are cheap) help to save labeled data (which are expensive). In the framework of *active learning*, where the learner selects some data from an unlabeled random sample and may ask for their labels, the saving of labels can be quite significant [10,11,1,3,12,14,4]. The picture is less clear in the framework of *semi-supervised learning* where the learner gets randomly chosen unlabeled and randomly chosen labeled data but has no influence on the selection of the latter. In this framework, it is usually assumed that there is a kind of compatibility between the target concept and the domain distribution.[2] There is a common intuition that, without assumptions of this kind, the benefit of semi-supervised learning is marginal. This intuition is supported by [5] whose authors consider the PAC-learning model without any extra-assumptions and compare a learner with full prior knowledge of the domain distribution D (representing the semi-supervised learner) with a learner that has no prior knowledge about D (the classical PAC-learner). They consider some basic concept classes over the real line and show that for absolutely continuous distributions D knowledge of D does not lower the label complexity by more than a constant factor. The main result in this paper goes in a similar direction but is incomparable to the findings in [5]: on one hand it is quite general by dealing with *arbitrary* finite concept classes; on the other hand, the higher level of generality comes at the price of introducing some limitations (with Limitation 2 being particularly annoying).

Relation of our Work to PAC-learning under a Fixed Distribution: Our results are also weakly related to [6] 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. 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 best upper and lower bounds known for PAC-learning under a fixed distribution 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.

[2] See the introduction of [9] for a discussion of the most popular assumptions, and see [2] for a systematic study of a PAC-like learning model augmented by compatibility-assumptions.

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, although it does not at all depend on D, does not perform much worse than the best strategy that is specialized to D. Section 4 is devoted to the general PAC-learning framework where the learner has no prior knowledge of the underlying domain distribution. It is shown, again by game-theoretic methods, that "Smart PAC-learners" do exist. Section 5 mentions some open problems.

2 Notations

We assume that the reader is familiar with the PAC-learning framework and knows the Minimax Theorem from game-theory [15].

Throughout the paper, we use the following notation:

- For every $n \geq 1$, let $[n] := \{1, \ldots .n\}$.
- X denotes a finite domain.
- $\mathcal{H} = \{h_1, \ldots, h_N\}$ denotes a finite concept class over domain X. Thus, every $h \in \mathcal{H}$ is a function of the form $h : X \rightarrow \{0, 1\}$. We shall assume henceforth that the hypothesis class coincides with $\mathcal{H}$.
- 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

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

  i.e., it maps labeled samples to hypotheses. $\mathcal{L}_1, \ldots, \mathcal{L}_M$ denotes the list of all learning functions.
- 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 $D(h \oplus h') \le \varepsilon$.

As usual, a hypothesis h is said to be *consistent with sample* (x, b) if $h(x) = b$. The *version space* for sample (x, b) is the set of all hypotheses from $\mathcal{H}$ being consistent with (x, b). A learning function $\mathcal{L}$ is said to be *consistent* if it maps every sample to a hypothesis from the corresponding version space.

A deterministic learner can be identified with a learning function (if computational issues are ignored). We consider, however, randomized learners. Each-one of these can be identified with a probability distribution over the set of all learning functions. A learner (or her strategy) is called *consistent* if she puts probability mass 1 on consistent learning functions.

3 Learning Under a Fixed Distribution Revisited

In this section, the domain distribution D is fixed and known to the learner. We shall describe PAC-learning under distribution D as a zero-sum game between two players: the learner who makes the first draw and her opponent who makes the second draw. This will offer the opportunity to apply the Minimax Theorem and to arrive at the "dual game" with the opponent making the first draw. Details follow.

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.,

$$I_D^{\varepsilon,x,b}[i, j] = \begin{cases} 1 \text{ if } D(\mathcal{L}_i(x, b) \oplus h_j) > \varepsilon \\ 0 \text{ otherwise} \end{cases} \quad . \tag{2}$$

Now, let

$$A_D^{\varepsilon,m}[i, j] := \mathbb{E}_{x \in D^m}\left[I_D^{\varepsilon,x,h_j(x)}[i, j]\right] = \sum_x D^m(x) I_D^{\varepsilon,x,h_j(x)}[i, j] \tag{3}$$

$$= \Pr_{x \sim D^m}\left[\mathcal{L}_i(x, h_j(x)) \text{ is } \varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D\right] \quad . \tag{4}$$

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 $\sum_{j=1}^N q_j = 1$). In the sequel, A_j denotes the j-th column of A. If the learners plays strategy p and her opponent plays strategy q (or the pure strategy j, resp.), the former has to pay an amount of $p^\top Aq$ (or an amount of $p^\top A_j$, resp.) to the latter.

This game models PAC-learning under distribution D in the sense that the following holds for given parameters m, ε, δ: there is a probabilistic learning strategy (= distribution over the learning functions that map a labeled sample of size m to a hypothesis) that, regardless of the choice of the target concept, leads to an ε-accurate hypothesis with a probability $1-\delta$ (or more) of success iff the row-player in the game with payoff-matrix $A = A_D^{\varepsilon,m}$ has a mixed strategy p such that for every pure strategy j of the opponent (i.e., for every choice of the target concept) $p^\top A_j \le \delta$.

We now put the Minimax Theorem in position. It states that

$$\min_p \max_q p^\top Aq = \max_q \min_p p^\top Aq \ . \tag{5}$$

In the sequel, we denote the optimal value in (5) by $\delta_D^*(\varepsilon, m)$. In order to establish a relation between $\delta_D^*(\varepsilon, m)$ and strategies for learners without prior knowledge of D (which will prepare the ground for the design of a smart PAC-learner in Section 4), we proceed as follows:

- We switch to the dual game with the opponent making the first draw.
- We consider a slightly modified dual game where the opponent makes the first draw, a labeled sample of size m is drawn at random afterwards, and finally the learner picks a hypothesis (pure strategy) or a distribution over hypotheses (mixed strategy).
- We describe a good strategy in the modified dual game that can be played without prior knowledge of D.

Consider the dual game with the opponent drawing first. 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 over $\mathcal{H}$ 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
     $$Q(j|x,b) = \begin{cases} \frac{q_j}{Q(b|x)} & \text{if } h_j(x) = b \\ 0 & \text{otherwise} \end{cases} \tag{6}$$
     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
$$\Pr(x,b) = D^m(x)Q(b|x) \ . \tag{7}$$

(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 $\tilde{A}$. We say that the learner's strategy $\tilde{p}$ is *consistent* if, for every (x,b), $\tilde{p}(x,b)$ puts probability mass 1 on the version space for (x,b).

(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 the following holds:*

(1) 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
$$p^\top Aq = \sum_{x,b} \Pr(x,b)\tilde{p}(x,b)^\top \tilde{A}\tilde{q}(x,b) \ . \tag{8}$$
Moreover, $\tilde{p}$ is a consistent strategy for the $\tilde{A}$-game if p is a consistent strategy for the A-game.

(2) The mapping $p \mapsto \tilde{p}$ is surjective, i.e., every mixed strategy for the learner in the $\tilde{A}$-game has a pre-image. Moreover, one can always find a consistent *strategy p as a pre-image of a* consistent *strategy $\tilde{p}$.*

PROOF. Recall that M is the number of learning functions of the form $\mathcal{L} : X^m \times \{0,1\}^m \rightarrow \mathcal{H}$. Thus, every probability vector $p \in [0,1]^M$ is a probability measure on the discrete product space $\Omega = \mathcal{H} \times \cdots \times \mathcal{H}$ with one factor $\mathcal{H}$ for every labeled sample $(x, b) \in X^m \times \{0,1\}^m$. Recall that $\mathcal{H} = \{h_1, \ldots, h_N\}$. Thus, every probability vector $\tilde{p}(x,b) \in [0,1]^N$ is a probability measure on the discrete space $\mathcal{H}$ which can be identified with factor (x, b) of Ω. We define the mapping $p \mapsto \tilde{p}$ by setting $\tilde{p}(x, b)$ equal to the marginal measure obtained by restricting p to factor (x, b) of Ω, i.e.,

$$\tilde{p}_{i'}(x,b) = \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} p_i \tag{9}$$

The following computation verifies (8):

$$\begin{aligned}
p^\top Aq &\overset{(3)}{=} \sum_x D^m(x) \sum_{j=1}^{N} \sum_{i=1}^{M} I^{x,h_j(x)}[i,j] p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i=1}^{M} I^{x,b}[i,j] p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i'=1}^{N} \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} \underbrace{I^{x,b}[i,j]}_{=\tilde{A}[i',j]} p_i q_j \\
&= \sum_{x,b} D^m(x) \sum_{j:h_j(x)=b} \sum_{i'=1}^{N} \tilde{A}[i',j] q_j \sum_{i:\mathcal{L}_i(x,b)=h_{i'}} p_i \\
&\overset{(9)}{=} \sum_{x,b} D^m(x) \sum_{i'=1}^{N} \sum_{j:h_j(x)=b} \tilde{A}[i',j] \tilde{p}_{i'}(x,b) q_j \\
&\overset{(6)}{=} \sum_{x,b} D^m(x) Q(b|x) \sum_{i'=1}^{N} \sum_{j=1}^{N} \tilde{A}[i',j] \tilde{p}_{i'}(x,b) Q(j|x,b) \\
&\overset{(7)}{=} \sum_{x,b} \Pr(x,b) \tilde{p}(x,b)^\top \tilde{A} \tilde{q}(x,b)
\end{aligned}$$

In order to show that $p \mapsto \tilde{p}$ is surjective, assume that a measure $\tilde{p}(x, y)$ on $\mathcal{H}$ is given for every labeled sample (x, b) and choose p as the corresponding product measure so that, for every $i = 1, \ldots, N$,

$$p_i = \prod_{x,b} \tilde{p}_{\mathcal{L}_i(x,b)}(x,b) \ . \tag{10}$$

Clearly, the marginal measure obtained by restricting p to factor (x, b) of Ω coincides with the measure $\tilde{p}(x, b)$ we started with. In other words, the product measure is a pre-image of $\tilde{p}$.

As far as consistency is concerned, finally note the following. If p puts probability mass 1 on consistent learning functions, then $\tilde{p}$, defined according to (9), puts probability mass 1 on the version space for (x, b). Conversely, if $\tilde{p}$ puts probability mass 1 on the version space for (x, b), then p, defined according to (10), puts probability mass 1 on consistent learning functions. □

In the sequel, we list some consequences of Lemma 1. For instance, the lemma immediately implies that the optimal value in the A-game (where $A = A_D^{\varepsilon,m}$) coincides with the optimal value in the $\tilde{A}$-game (where $\tilde{A} = \tilde{A}_D^{\varepsilon}$), i.e.,

$$\delta_D^*(\varepsilon, m) = \min_p \max_q p^\top A q = \max_q \left[\sum_{x,b} \left(\Pr(x, b) \cdot \min_{\tilde{p}(x,y)} \left[\tilde{p}(x, y)^\top \tilde{A} \tilde{q}(x, b) \right] \right) \right] . \tag{11}$$

Remark 2 *Notice that (11) offers the opportunity to prove (non-constructively) the* existence *of a good learning strategy for the A-game with the learner making the first draw, by presenting a good (sample-dependent) learning strategy for the $\tilde{A}$-game with the learner making the second draw.*

The next two results concern "trivial domain distributions" whose prior knowledge leads to zero-loss in the A-game.

Corollary 3 *If $\delta_D^*(\varepsilon, m) = 0$, then the following holds: for every $x \in X^m$ such that $D^m(x) > 0$, and for every $b \in \{0, 1\}^m$, there exists a hypothesis $h^* \in \mathcal{H}$ that is ε-accurate for every hypothesis in the version space for (x, b).*

PROOF. Assume that there exists an $x \in X^m$ such that $D^m(x) > 0$, and there exists a $b \in \{0, 1\}^m$ such that

$$\forall h^* \in \mathcal{H}, \exists h \in \mathcal{H} : \ h(x) = b \ \wedge \ D(h^* \oplus h) > \varepsilon \ . \tag{12}$$

We have to show that this assumption leads to the conclusion $\delta_D^*(\varepsilon, m) > 0$. To this end, pick x and b such that $D^m(x) > 0$ and such that (12) is valid. Let q be the uniform distribution on $\mathcal{H}$. According to (7), we may conclude that $\Pr(x, b) > 0$. An inspection of (11) now reveals that $\delta_D^*(\varepsilon, m) > 0$, as desired. □

Corollary 4 *If $\delta_D^*(\varepsilon, m) = 0$, then every consistent learner suffers loss 0 in the $A_D^{2\varepsilon,m}$-game.*

PROOF. According to Lemma 1, it suffices to show that every strategy $\tilde{p}(x, b)$ for the $\tilde{A}$-game that puts probability mass 1 on the version space for (x, b) suffers loss 0 in the $\tilde{A}$-game. According to Corollary 3, there exists a hypothesis h^* that is ε-accurate for every hypothesis in the version space V for (x, b). This clearly implies, that every hypothesis in V is 2ε-accurate for every other function in V. Thus, putting probability mass 1 on V leads to loss 0. □

We close this section with a result that prepares the ground for our analysis of general PAC-learners in Section 4:

Lemma 5 *Let $\varepsilon > 0$ be a given accuracy, and let $m \geq 1$ be a given sample size. For every probability vector $q \in [0,1]^N$, and every domain distribution D, the following holds:*

$$\sum_{x,b} \Pr(x,b)\tilde{q}(x,b)^\top \tilde{A}_D^{2\varepsilon}\tilde{q}(x,b) \leq 2\delta_D^*(\varepsilon,m) \tag{13}$$

PROOF. The left hand-side in (13) represents the learners loss in the $\tilde{A}_D^{2\varepsilon}$-game when the opponent plays strategy q (the à-priori probabilities for the target concepts) and the learner plays strategy $\tilde{q}(x,b)$ on sample (x,b). Recall that $\tilde{q}(x,b)$ is the collection of à-posteriori probabilities for the target concepts. 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 learner's loss. 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. Given q and (x,b), the term $\tilde{A}_i^\varepsilon\tilde{q}(x,b)$ represents the loss suffered by a learner with hypothesis h_i in the $\tilde{A}_D^\varepsilon$-game. This loss equals the total à-posteriori-probability of the hypotheses from the version space that are *not* ε-close to h_i w.r.t. domain distribution D. It is minimized by picking 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. With this notation, it follows that

$$\sum_{x,b} \Pr(x,b)\tilde{A}_{i^*(x,b)}^\varepsilon\tilde{q}(x,b) \leq \delta_D^*(\varepsilon,m) \tag{14}$$

with equality if the strategy q of the learner's opponent is optimal. The situation is visualized in Figure 1. We are now prepared to verify (13). Assume that, given q and (x,b), the learner applies strategy $\tilde{q}(x,b)$ instead of choosing the best hypothesis h^*. There are two "unlucky cases":

- The learners random hypothesis falls into the hatched area in Figure 1.
- The opponent's random target concept falls into this hatched area.

Each of the unlucky events happens with a probability that equals the total à-posteriori probability of the hypotheses in the hatched area, i.e. it happens with probability $\tilde{A}_{i^*(x,b)}^\varepsilon\tilde{q}(x,b)$, respectively. If none of the unlucky events occurs, then the learner's hypothesis *and* the target concept fall into the circle

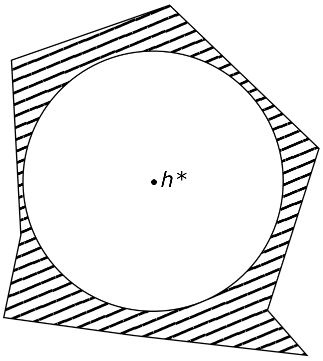

Fig. 1. The polygon represents the version space for (x, b). The circle represents the hypotheses that are ε-close to h^* w.r.t. D. The hatched area represents the loss induced by h^* on sample (x, b).

in Figure 1 so that they are 2ε-close to each other w.r.t. D. Our discussion shows that

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

which, in view of (14), yields (13). □

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 5, this is a reasonably good strategy for *any* fixed underlying domain distribution. Note furthermore that a learner with strategy $\tilde{q}(x, b)$ is consistent, since the à-posteriori probabilities assign probability mass 1 to the version space for (x, b).

The puzzled reader might ask why we are not done with our design of a smart PAC-learner: we ended-up with a good strategy for the learner that can be played without prior knowledge of the domain distribution. So what? The problem is, however, that so far we considered a game that models PAC-learning under a fixed distribution. A game that models general PAC-learning would correspond to a payoff-matrix that decomposes into blocks with one block per domain distribution.[3] It turns out, unfortunately, that this considerably increases the power of the opponent. In particular, he could play a mixed strategy that fixes

- a probability measure μ on the $|X|$-dimensional probability simplex

$$\Delta := \{z \in [0,1]^{|X|} :\ z_1 + \cdots + z_{|X|} = 1\} \ , \tag{15}$$

  where every $z \in \Delta$ represents a measure D_z putting probability mass z_ν on the ν-th element from X for $\nu = 1, \ldots, |X|$,

[3] Ignore for the moment the fact that there are infinitely many blocks. That wouldn't pose a problem because the Minimax Theorem applies even when the pure strategies form an infinite but compact set.

- *conditional* à-priori probabilities $q(j|z)$ for choosing target concept h_j provided that D_z is the underlying domain distribution.

In this general setting, the à-posteriori probabilities, associated with the hypotheses after having seen a labeled sample (x, b), previously denoted $\tilde{q}(x, b)$, are now conditioned on z as well and therefore denoted $\tilde{q}(x, b|z)$. Consequently, the loss suffered by a learner playing strategy $\tilde{p}$ formally looks as follows (compare with (11)):

$$\mathbb{E}_{z\sim\mu}\left[\sum_{x,b}\Pr(x, b|z)\tilde{p}(x, b)^\top \tilde{A}^{2\varepsilon}_{D_z}\tilde{q}(x, b|z)\right] \tag{16}$$

According to Lemma 5, the learner can favorably play strategy $\tilde{p} = \tilde{q}$ in the fixed-distribution setting. As can be seen from (16), there is no uniform good choice for $\tilde{p}$ anymore in the general setting since $\tilde{q}$ is conditioned to z. For this reason, we are not done yet and move on to Section 4.

4 Smart PAC-Learners

Throughout this section we set $d := |X|$ and $X := \{\xi_1, \ldots, \xi_d\}$. As specified in (15), Δ denotes the d-dimensional probability simplex so that, for every $z \in \Delta$ and for all $\nu = 1, \ldots, d$, $D_z(\xi_\nu) = z_\nu$.

Ideally, we would like to design a smart PAC-learner who performs well in comparison to a learner with full prior knowledge of D_z for *all choices* of $z \in \Delta$. This (somewhat overambitious) endeavor is briefly described (in terms of an appropriately defined zero-sum game) in Section 4.1. After a short discussion of some measurability issues in Section 4.2, we pursue a less ambitious goal in Section 4.3 (modeled again as a zero-sum game between the learner and her opponent) and show that there is a smart PAC-learner who performs well in comparison to a learner with full prior knowledge of D_z for *most choices* of $z \in \Delta$. The total probability of the region containing the "bad choices" of z can be made as small as we like (at the expense of a slightly degraded performance on the "good choices"). Moreover, the underlying probability measure on Δ can be specified by a "user" (who can try to put high probability mass on realistic measures), and the strategy of the smart PAC-learner can be chosen such as to adapt to this specification.

Before dipping into technical details, we would like to provide the reader with a survey on the various zero-sum games (each-one given by a payoff-matrix) which are relevant in the sequel:

(1) The basic building stone is the matrix $A = A^{\varepsilon,m}_D$. The corresponding

zero-sum game models PAC-learning under the fixed distribution D.

(2) When switching to classical PAC-learning with arbitrary distributions, it looks natural to analyze a block-matrix that reserves one block $A_{D_z}^{\varepsilon,m}$ for every $z \in \Delta$. It will however be more clever to insert a scaling factor $1/\delta_z^*$ in block z where $\delta_z^* = \delta_{D_z}^*(\varepsilon, m)$ is the best possible expected failure-rate of a learner with full prior knowledge of D_z: this will force the learner to choose a strategy that achieves a small ratio between her own expected failure-rate and δ_z^*. The payoff-matrix obtained after scaling is denoted R. A learner playing successfully the R-game would achieve a small "worst performance ratio".

(3) Since we were not able to find a good strategy for the learner in the R-game, we switch to the $\bar{R}$-game. The payoff-matrix $\bar{R}$ results from R basically by averaging over all $z \in \Delta$. A learner playing successfully the $\bar{R}$-game achieves a small "average performance ratio". We shall find a good strategy for the learner in the $\bar{R}$-game, and this will finally lead us to our main result.

4.1 Towards a Small Worst Performance Ratio

In this section, we consider a game with a payoff-matrix R that is defined blockwise so that, for every $z \in \Delta$, $R^{(z)}$ denotes the block reserved for distribution D_z. Every block has M rows (one row for every learning function) and N columns (one column for every possible target concept). Before we present the definition of $R^{(z)}$, we fix some notation.

Recall from the previous section that

$$A_{D_z}^{2\varepsilon,m}[i,j] = \Pr_{x \sim D_z^m}[\mathcal{L}_i(x, h_j(x)) \text{ is } 2\varepsilon\text{-inaccurate for } h_j \text{ w.r.t. } D_z] \enspace .$$

When parameters ε, m are obvious from context, we shall simply write $A^{(z)}$ instead of $A_{D_z}^{2\varepsilon,m}$. The j'th column of this matrix is then written as $A_j^{(z)}$. With this notation, the following holds: if the learner chooses learning function $\mathcal{L}_i$ with probability p_i, she will fail to deliver a 2ε-accurate hypothesis with probability

$$\max_{(z,j) \in \Delta \times [N]} p^\top A_j^{(z)}$$

in the worst-case. Recall that $\delta_z^* = \delta_{D_z}^*(\varepsilon, m)$ denotes the best possible expected failure-rate of a learner with full prior knowledge of D_z. Thus,

$$\frac{\max_{j \in [N]} p^\top A_j^{(z)}}{\delta_z^*}$$

measures how well the PAC-learner with "strategy" p performs in relation to the best learner with full prior knowledge of D_z. It is therefore reasonable to choose the block-matrix $R^{(z)}$ as follows:

$$R^{(z)}[i,j] := \begin{cases} \frac{1}{\delta^*_{D_z}(\varepsilon,m)} \cdot A^{2\varepsilon,m}_{D_z}[i,j] & \text{if } \delta^*_{D_z}(\varepsilon,m) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

With the convention $0/0 = 1$, the quantity

$$\rho^p(\varepsilon,m) := \max_{(z,j)\in\Delta\times[N]} p^\top R^{(z)}_j = \max_{(z,j)\in\Delta\times[N]} \frac{p^\top A^{(z)}_j}{\delta^*_z}$$

is called the *worst performance ratio* of strategy p.[4] The value

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

is the best possible worst performance ratio. Note that it coincides with the smallest loss suffered by the learner in the game with payoff-matrix R. A very strong result would be $\rho^*(\varepsilon,m) \le c$ for some small constant c. But, since this is somewhat overambitious, we shall pursue a weaker goal in the following.

4.2 Some Measurability Issues

Before we turn our attention to smart PAC-learners, we have to clarify first some measurability issues. Let $(\Omega, \mathcal{A}, \mu)$ be an arbitrary measure space, and let $f(\omega)$ be a numerical function in variable $\omega \in \Omega$. The following facts are well-known:

- f is measurable if and only if

$$\forall \alpha \in \mathbb{R} : \{\omega \in \Omega :\ f(\omega) \ge \alpha\} \in \mathcal{A} \ . \tag{17}$$

- The sum and the product of two measurable functions is measurable.
- The reciprocal of a strictly positive (or strictly negative, resp.) measurable function is measurable.
- The infimum (or supremum, resp.) of a sequence of measurable functions is measurable.
- If Ω is a Borel set equipped with the Borel-algebra and $f(\omega)$ is a continuous function in $\omega \in \Omega$, then f is measurable.

These closure properties have the following implications:

[4] Note that, according to Corollary 4, $\delta^*_z = 0$ implies that, for every consistent strategy p, $p^\top A^{(z)}_j/\delta^*_z = 0/0 = 1$.

Corollary 6 *Let Ω be a Borel set equipped with the Borel-algebra. Let K be a finite or countably infinite set and, for every $k \in K$, let Ω_k be a Borel-set such that $\Omega = \cup_{k \in K} \Omega_k$. Let $f(\omega)$ be a numerical function in $\omega \in \Omega$ that is continuous on every Ω_k. Then f is measurable.*

PROOF. For every $\alpha \in \mathbb{R}$, consider the decomposition

$$\{\omega \in \Omega | \ f(\omega) \geq \alpha\} = \bigcup_{k \in K} \{\omega \in \Omega_k | \ f(\omega) \geq \alpha\} \ .$$

Since f is continuous on every Ω_k, the right hand-side is a countable union of Borel-sets and therefore a Borel-set itself. Thus, f satisfies (17) and is a measurable function. □

Corollary 7 *Let the d-dimensional probability simplex Δ be equipped with the Borel-algebra. Then, for every choice of $\varepsilon, m, x, b, i, j$, $I_{D_z}^{\varepsilon,x,b}[i,j]$, $A_{D_z}^{\varepsilon,m}[i,j]$, and δ_z^* are measurable functions in $z \in \Delta$.*

PROOF. Since

$$A_{D_z}^{\varepsilon,m}[i,j] = \sum_x D_z^m(x) I_{D_z}^{\varepsilon,x,h_j(x)}[i,j] \text{ and } \delta_z^* = \min_p \max_q p^\top A_{D_z}^{\varepsilon,m} q \ ,$$

it suffices to show that $I_{D_z}^{\varepsilon,x,b}[i,j]$ is a measurable function in $z \in \Delta$. For any $T \subseteq [d]$, consider the hyperplane

$$H_T := \left\{ z : \ \sum_{\nu \in T} z_\nu = \varepsilon \right\}$$

and the halfspaces

$$H_T^+ := \left\{ z \in \Delta : \ \sum_{\nu \in T} z_\nu > \varepsilon \right\} \ , \ H_T^- := \left\{ z \in \Delta : \ \sum_{\nu \in T} z_\nu \leq \varepsilon \right\} \ .$$

These halfspaces decompose Δ into finitely many cells where every single cell can be written as an intersection of Δ with finitely many halfspaces (which clearly yields a Borel-set). An inspection of (2) shows that the discontinuities of $I_{D_z}^{\varepsilon,x,b}[i,j]$ occur on the hyperplane

$$H_T \text{ such that } T = \{\nu \in [d] : \ \mathcal{L}_i(x,b)(\xi_\nu) \neq h_j(\xi_\nu)\}$$

only and that $I_{D_z}^{\varepsilon,x,b}[i,j]$ is continuous on every single cell of the mentioned decomposition. According to Corollary 6, $I_{D_z}^{\varepsilon,x,b}[i,j]$ is a measurable function in $z \in \Delta$. □

Let μ denote an arbitrary but fixed probability measure on Δ w.r.t. the algebra of d-dimensional Borel-sets. For every $\zeta > 0$, consider the following decomposition of Δ:

$$\begin{aligned} \Delta_0 &= \{z \in \Delta : \ \delta_z^* = 0\} \\ \Delta_1(\zeta) &= \{z \in \Delta : \ 0 < \delta_z^* < \zeta\} \\ \Delta_2(\zeta) &= \{z \in \Delta : \ \delta_z^* \geq \zeta\} \end{aligned}$$

Note that these sets are Borel-sets since δ_z^* is a measurable function according to Corollary 7. Since probability measures are continuous from above, we get

$$\lim_{\zeta \to 0} \mu(\Delta_1(\zeta)) = \mu\left(\bigcap_{\zeta > 0} \Delta_1(\zeta)\right) = \mu(\emptyset) = 0 \ . \tag{18}$$

In our discussion of smart PAC-learners, it is justified to focus on domain distributions D_z such that $z \in \Delta_2(\zeta)$ for the following reasons:

- Distributions D_z such that $z \in \Delta_0$ do not pose any problem to a consistent learner. Compare with Corollary 4.
- Distributions D_z such that $z \in \Delta_1(\zeta)$ might pose a problem to PAC-learners but the probability mass assigned to them by μ can be made arbitrarily small according to (18).

Let μ' be the probability measure induced by μ on $\Delta_2(\zeta)$. Consider the following $M \times N$ payoff-matrix:

$$\bar{R}^\zeta[i,j] := \mathbb{E}_{z \sim \mu'}\left[R^{(z)}[i,j]\right] = \frac{1}{\mu(\Delta_2(\zeta))} \cdot \int_{\Delta_2(\zeta)} R^{(z)}[i,j] \, d\mu$$

Note that the integral exists because, for every $z \in \Delta_2(\zeta)$, $\delta_{D_z}^*(\varepsilon, m) \geq \zeta$ so that

$$R^{(z)}[i,j] = \frac{1}{\delta_{D_z}^*(\varepsilon, m)} \cdot A_{D_z}^{2\varepsilon,m}[i,j]$$

is bounded by $1/\zeta$ and measurable according to Corollary 7. The quantity

$$\bar{\rho}^p(\varepsilon, m) := \max_{j \in [N]} p^\top \bar{R}_j^\zeta$$

is called the *average performance ratio* of strategy p (where the target concept is still chosen in a worst-case-fashion but the domain distribution is chosen at random according to μ'). According to the Minimax Theorem, the following holds:

$$\min_p \max_q p^\top \bar{R}^\zeta q = \max_q \min_p p^\top \bar{R}^\zeta q \tag{19}$$

Clearly, the optimal value in (19) coincides with the best possible average performance ratio. Equation (19) offers the opportunity to (non-constructively) show the *existence* of a "good" learning strategy with the learner making the first draw, by presenting a "good" learning strategy with the learner making the second draw, where "good" here means "achieving a small average performance ratio". (Compare with Remark 2.)

Lemma 8 *For every mixed strategy q of the opponent in the $\bar{R}^\zeta$-game, there exists a consistent mixed strategy p for the learner such that $p^\top \bar{R}^\zeta q \le 2$.*

PROOF. From the decomposition (8), we get the following decomposition of $p^\top \bar{R}^\zeta q$:

$$\begin{aligned} p^\top \bar{R}^\zeta q &= p^\top \left(\mathbb{E}_{z\sim\mu'}\left[\bar{R}^{(z)}\right]\right) q \\ &= \mathbb{E}_{z\sim\mu'}\left[\frac{1}{\delta_z^*} p^\top A^{(z)} q\right] = \mathbb{E}_{z\sim\mu'}\left[\frac{1}{\delta_z^*}\sum_{x,b}\Pr(x,b|z)\tilde{p}(x,b)\tilde{A}^{(z)}\tilde{q}(x,b)\right] \end{aligned}$$

Here, $\delta_z^* = \delta_{D_z}^*(\varepsilon, m)$, $\Pr(x,b|z) = D_z^m(x)Q(b|x)$, $\tilde{A}^{(z)} = \tilde{A}_{D_z}^{2\varepsilon}$, and the quantities $Q(b|x), \tilde{q}, \tilde{p}$ are derived from q and p, respectively, as explained in Section 3. According to Lemma 5 (applied to $D = D_z$),

$$\sum_{x,b}\Pr(x,b|z)\tilde{q}(x,b)\tilde{A}^{(z)}\tilde{q}(x,b) \le 2\delta_z^* \ . \tag{20}$$

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

$$p^\top \bar{R}^\zeta q = \mathbb{E}_{z\sim\mu'}\left[\frac{1}{\delta_z^*}\sum_{x,b}\Pr(x,b|z)\tilde{p}(x,b)\tilde{A}^{(z)}\tilde{q}(x,b)\right] \overset{(20)}{\le} 2 \ ,$$

as desired. □

Corollary 9 *For every probability measure μ on Δ, there exists a consistent mixed strategy for the learner with an average performance ratio of at most* 2.

PROOF. The Minimax Theorem (applied to the submatrix of $\bar{R}^\zeta$ with rows corresponding to consistent learning functions) combined with Lemma 8 yields the result. □

Corollary 9 talks about the average performance ratio which, by definition, is the performance ratio averaged over $\Delta_2(\zeta)$ only. Furthermore, it does not explicitly bound the probability mass of domain distributions D_z for which the smart PAC-learner performs considerably worse than the learner with full prior knowledge of D_z. The next result, the main result in this paper, fills these gaps:

Theorem 10 *For every probability measure μ on Δ and for every $c > 1$, $\gamma > 0$, there exists a mixed strategy p for the learner such that, for $j = 1, \ldots, N$,*

$$\mu\left(\left\{z \in \Delta :\ p^\top R_j^{(z)} \geq 2c\right\}\right) < \frac{1}{c} + \gamma\ .$$

PROOF. For sake of brevity, let $E := \{z \in \Delta :\ p^\top R_j^{(z)} \geq 2c\}$. With this notation, $\mu(E)$ is bounded from above by

$$\mu\left(E|\Delta_0\right) \cdot \mu(\Delta_0) + \mu\left(E|\Delta_1(\zeta)\right) \cdot \mu(\Delta_1(\zeta)) + \mu\left(E|\Delta_2(\zeta)\right) \cdot \mu(\Delta_2(\zeta))\ .$$

The first term contributes 0 because of Corollary 4. The second-one contributes at most $\mu(\Delta_1(\zeta))$ which is smaller than γ for every sufficiently small ζ. The third-one contributes at most $1/c$ according to Corollary 9 combined with Markov's inequality. Thus, $\Pr_{z \sim \mu}[E] < 1/c + \gamma$, as desired. □

According to Theorem 10, 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 μ) not belonging to the "vast majority" is bounded by $1/c + \gamma$ (where γ may be chosen as small as we like). So Theorem 10 is the result that we had announced in the introduction.

5 Open Problems

- For every finite hypothesis class, we have shown the mere existence of a smart PAC-learner 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 smart PAC-learner whose *worst* performance ratio is bounded by a small constant.
- In the present paper, we restricted ourselves to finite domains X for ease of technical exposition (e.g., compactness of the $|X|$-dimensional probability simplex). We conjecture however that this restriction can be weakened considerably.

- In the present paper, we restricted ourselves to the *non-agnostic* setting. Explore whether there are smart PAC-learners in the *agnostic* setting.
- In the present paper, we restricted ourselves to the PAC-learning model without adding any extra-assumptions concerning the compatibility between the target concept and the domain distribution (as they are typically made in the framework of semi-supervised learning). Clarify which kind of extra-assumptions gives significantly more advantage to semi-supervised learners than to (smart) fully supervised-ones.

As for the last open question, we follow a suggestion of one of the referees and add a little comment. It is well-known that in the Co-training Model of Learning under the Conditional Independence Assumption [7], essentially one labeled example is sufficient for a semi-supervised learner provided there are sufficiently many unlabeled random examples (e.g., see Theorem 15 in [2]). For the same model, we can prove rigorously that every fully supervised learner asymptotically requires at least $\Omega\left(\sqrt{\frac{d_1 \cdot d_2}{\varepsilon}}\right)$ labeled examples (and there exist classes for which this bound is tight up to logarithmic factors). Here, d_1 and d_2 denote the VC-dimensions of the two concept classes to be learned in the Co-training model. Thus, as correctly predicted by one of the referees, the last open problem in our list is partially solved: the Co-training Model of Learning under the Conditional Independence Assumption indeed gives significantly more advantage to semi-supervised learners than to the best fully supervised-one. [5] But the issue certainly needs further clarification.

Acknowledgments: We would like to thank two unknown referees for their insightful and inspiring comments and suggestions.

References

[1] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[2] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the Association on Computing Machinery*, 57(3):19:1–19:46, 2010.

[3] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

[5] Note, however, that sample size bound $\tilde{O}(\sqrt{d_1 \cdot d_2/\varepsilon})$, occasionally achievable for fully supervised learners in this model, breaks the barrier $\Omega(d/\varepsilon)$ for PAC-learning without co-training.

[4] Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active lerning. *Machine Learning*, 80(2–3):111–139, 2010.

[5] 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.

[6] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[7] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[8] 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.

[9] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning.* MIT Press, 2006.

[10] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(3):201–221, 1994.

[11] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, pages 235–242. MIT Press, 2006.

[12] Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, pages 353–360. MIT Press, 2008.

[13] 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.

[14] Steve Hanneke. A bound on the label complexity of active learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 353–360, 2007.

[15] Peter Morris. *Introduction to Game Theory.* Springer, 1994.