Supervised Learning and Co-training ⋆

Malte Darnstädt[1], Hans Ulrich Simon[1], and Balázs Szörényi[2]

[1] Fakultät für Mathematik, Ruhr-Universität Bochum
D-44780 Bochum, Germany
{malte.darnstaedt,hans.simon}@rub.de

[2] Hungarian Academy of Sciences and University of Szeged
Research Group on Artificial Intelligence
H-6720 Szeged, Hungary
szorenyi@inf.u-szeged.hu

Abstract. Co-training under the Conditional Independence Assumption is among the models which demonstrate how radically the need for labeled data can be reduced if a huge amount of unlabeled data is available. In this paper, we explore how much credit for this saving must be assigned solely to the extra-assumptions underlying the Co-training model. To this end, we compute general (almost tight) upper and lower bounds on the sample size needed to achieve the success criterion of PAC-learning within the model of Co-training under the Conditional Independence Assumption in a purely supervised setting. The upper bounds lie significantly below the lower bounds for PAC-learning without Co-training. Thus, Co-training saves labeled data even when not combined with unlabeled data. On the other hand, the saving is much less radical than the known savings in the semi-supervised setting.

1 Introduction

In the framework of semi-supervised learning, it is usually assumed that there is a kind of compatibility between the target concept and the domain distribution.[3] This intuition is supported by recent results indicating that, without extra-assumptions, there exist purely supervised learning strategies which can compete fairly well against semi-supervised learners (or even against learners with full prior knowledge of the domain distribution) [3, 7].

In this paper, we go one step further and consider the following general question: given a particular extra-assumption which makes semi-supervised learning quite effective, how much credit must be given to the extra-assumption alone? In other words, to which extent can labeled examples be saved by exploiting the extra-assumption in a purely supervised setting? We provide a first answer to this question in a case study which is concerned with the model of Co-training under the Conditional Independence Assumption [4]. In this model (whose formal definition will be recalled in Section 2), *one labeled example* is enough for

⋆ This work was supported by the bilateral Research Support Programme between Germany (DAAD 50751924) and Hungary (MÖB 14440).

[3] See the introduction of [6] for a discussion of the most popular assumptions.

achieving the success criterion of PAC-learning provided that there are sufficiently many unlabeled examples [1].[4] Recall that PAC-learning without any extra-assumption requires d/ε labeled samples (up to logarithmic factors) where d denotes the VC-dimension of the concept class and ε is the accuracy parameter [5]. The step from d/ε to just a single labeled example is a giant-one. In this paper, we show however that part of the credit must be assigned to just the Co-training itself. More specifically, we show that the number of sample points needed to achieve the success criterion of PAC-learning in the purely supervised model of Co-training under the Conditional Independence Assumption has a linear growth in $\sqrt{d_1 d_2/\varepsilon}$ (up to some hidden logarithmic factors) as far as the dependence on ε and on the VC-dimensions of the two involved concept classes is concerned. Note that, as ε approaches 0, $\sqrt{d_1 d_2/\varepsilon}$ becomes much smaller than the well-known lower bound $\Omega(d/\varepsilon)$ on the number of examples needed by a traditional (not co-trained) PAC-learner.

The remainder of the paper is structured as follows. Section 2 clarifies the notations and formal definitions that are used throughout the paper and mentions some elementary facts. Section 3 presents a fundamental inequality that relates a suitably defined variant of Hanneke's disagreement coefficient [9] to a purely combinatorial parameter, $s(\mathcal{C})$, which is closely related to the "unique negative dimension" from [8]. This will later lead to the insight that the product of the VC-dimension of a (suitably chosen) hypothesis class and a (suitably defined) disagreement coefficient has the same order of magnitude as $s(\mathcal{C})$. Section 3 furthermore investigates how a concept class can be padded so as to increase the VC-dimension while keeping the disagreement coefficient invariant. The padding can be used to lift lower bounds that hold for classes of low VC-dimension to increased lower bounds that hold for some classes of arbitrarily large VC-dimension. The results of Section 3 seem to have implications for active learning and might be of independent interest. Section 4.1 presents some general upper bounds in terms of the relevant learning parameters (including ε, the VC-dimension, and the disagreement coefficient, where the product of the latter two can be replaced by the combinatorial parameters from Section 3). Section 4.2 shows that all general upper bounds from Section 4.1 are (nearly) tight. Interestingly, the learning strategy that is best from the perspective of a worstcase analysis has one-sided error. Section 4.3 presents improved bounds for classes with special properties. Section 5 contains some final remarks.

Due to space constraints, not all proofs are given in full detail. We plan to give the missing parts in an upcoming journal version of this paper.

[4] This is one of the results which impressively demonstrate the striking potential of properly designed semi-supervised learning strategies although the underlying compatibility assumptions are somewhat idealized and therefore not likely to be strictly satisfied in practice. See [2, 12] for suggestions of relaxed assumptions.

2 Definitions, Notations, and Facts

We assume the reader is familiar with Valiant's model of Probably Approximately Correct Learning (PAC-learning) [11]. In Co-training [4], it is assumed that there is a pair of concept classes, $\mathcal{C}_1$ and $\mathcal{C}_2$, and that random examples come in pairs $(x_1, x_2) \in X_1 \times X_2$. Moreover, the domain distribution D, according to which the random examples are generated, is perfectly compatible with the target concepts, say $h_1^* \in \mathcal{C}_1$ and $h_2^* \in \mathcal{C}_2$, in the sense that $h_1^*(x_1) = h_2^*(x_2)$ with probability 1. (For this reason, we sometimes denote the target label as $h^*(x_1, x_2)$.) As in [4, 1], our analysis builds on the Conditional Independence Assumption: x_1, x_2, considered as random variables that take "values" in X_1 and X_2, respectively, are conditionally independent given the label. As in [1], we perform a PAC-style analysis of Co-training under the Conditional Independence Assumption. But unlike [1], we assume that there is no access to unlabeled examples. The resulting model is henceforth referred to as the "PAC Co-training Model under the Conditional Independence Assumption".

Let $\mathcal{C}$ be a concept class over domain X and $\mathcal{H} \supseteq \mathcal{C}$ a hypothesis class over the same domain. For every $h^* \in \mathcal{C}$ and every $X' \subseteq X$, the corresponding *version space in* $\mathcal{H}$ is given by $V_{\mathcal{H}}(X', h^*) := \{h \in \mathcal{H} |\ \forall x \in X' : h(x) = h^*(x)\}$. Let $V \subseteq \mathcal{C}$. The *disagreement region of* V is given by $\mathrm{DIS}(V) := \{x \in X |\ \exists h, h' \in V : h(x) \neq h'(x)\}$. Let $\mathbb{P}$ denote a probability measure on X. We define the following variants of disagreement coefficients:

$$\theta(\mathcal{C}, \mathcal{H} | \mathbb{P}, X', h^*) := \frac{\mathbb{P}(\mathrm{DIS}(V_{\mathcal{C}}(X', h^*)))}{\sup_{h \in V_{\mathcal{H}}(X', h^*)} \mathbb{P}(h \neq h^*)}$$

$$\theta(\mathcal{C}, \mathcal{H}) := \sup_{\mathbb{P}, X', h^*} \theta(\mathcal{C}, \mathcal{H} | \mathbb{P}, X', h^*)$$

For sake of brevity, let $\theta(\mathcal{C}) := \theta(\mathcal{C}, \mathcal{C})$. Note that

$$\theta(\mathcal{C}, \mathcal{H}) \leq \theta(\mathcal{C}) \leq |\mathcal{C}| - 1 \ . \tag{1}$$

The first inequality is obvious from $\mathcal{C} \subseteq \mathcal{H}$ and $h^* \in \mathcal{C}$, the second follows from

$$\mathrm{DIS}(V_{\mathcal{C}}(X', h^*)) = \bigcup_{h \in V_{\mathcal{C}}(X', h^*) \setminus \{h^*\}} \{x |\ h(x) \neq h^*(x)\}$$

and an application of the union bound.

As an example we will calculate θ for the following class, which will also be useful for proving lower bounds in section 4.2:

$$\mathrm{SF}_n = \{\{0\}, \{0, 1\}, \{0, 2\}, \dots, \{0, n\}\}$$

Lemma 1. *$\theta(\mathrm{SF}_n) = n$.*

Proof. Let $\mathbb{P}$ be uniform on $\{1, \dots, n\}$, let $X' = h^* = \{0\}$. Then $V := V_{\mathrm{SF}_n}(X', h^*) = \mathrm{SF}_n$ and $\mathrm{DIS}(V) = \{1, \dots, n\}$ has probability mass 1. Thus,

$$\theta(\mathrm{SF}_n) \geq \theta(\mathrm{SF}_n, \mathrm{SF}_n | \mathbb{P}, X', h^*) = \frac{\mathbb{P}(\mathrm{DIS}(V))}{\sup_{h \in V} \mathbb{P}(h \neq h^*)} = \frac{1}{1/n} = n \ .$$

Conversely, $\theta(SF_n) \leq |\mathrm{SF}_n| - 1 = n$ (according to (1)). □

The main usage of this disagreement coefficient is as follows. First note that we have $\mathbb{P}(DIS(V_{\mathcal{C}}(X',h^*))) \leq \theta(\mathcal{C},\mathcal{H}) \cdot \sup_{h \in V_{\mathcal{H}}(X',h^*)} \mathbb{P}(h \neq h^*)$ for every choice of $\mathbb{P}, X', h^*$. This inequality holds in particular when X' consists of m points in X chosen independently at random according to $\mathbb{P}$. According to classical sample size bounds in PAC-learning, there exists a sample size $m = \tilde{O}(\mathrm{VCdim}(\mathcal{H})/\varepsilon)$ such that, with probability at least $1-\delta$, $\sup_{h \in V_{\mathcal{H}}(X',h^*)} \mathbb{P}(h \neq h^*) \leq \varepsilon$. Thus, with probability at least $1-\delta$ (taken over the random sample X'), $\mathbb{P}(DIS(V_{\mathcal{C}}(X',h^*))) \leq \theta(\mathcal{C},\mathcal{H}) \cdot \varepsilon$. This discussion (with $\varepsilon/\theta(\mathcal{C},\mathcal{H})$ substituted for ε) is summarized in the following

Lemma 2. *There exists a sample size $m = \tilde{O}(\theta(\mathcal{C},\mathcal{H}) \cdot \mathrm{VCdim}(\mathcal{H})/\varepsilon)$ such that the following holds for every probability measure $\mathbb{P}$ on domain X and for every target concept $h^* \in \mathcal{C}$. With probability $1-\delta$, taken over a random sample X' of size m, $\mathbb{P}((DIS(V_{\mathcal{C}}(X',h^*))) \leq \varepsilon$.*

This lemma indicates that one should choose $\mathcal{H}$ so as to minimize $\theta(\mathcal{C},\mathcal{H}) \cdot \mathrm{VCdim}(\mathcal{H})$. Note that making $\mathcal{H}$ more powerful leads to smaller values of $\theta(\mathcal{C},\mathcal{H})$ but comes at the prize of an increased VC-dimension.

We say that $\mathcal{H}$ *contains hypotheses with plus-sided errors (or minus-sided errors, resp.) w.r.t. concept class* $\mathcal{C}$ if, for every $X' \subseteq X$ and every $h^* \in \mathcal{C}$, there exists $h \in V_{\mathcal{H}}(X',h^*)$ such that $h(x) = 0$ ($h(x) = 1$, resp.) for every $x \in \mathrm{DIS}(V_{\mathcal{C}}(X',h^*))$. A sufficient (but, in general, not necessary) condition for a class $\mathcal{H}$ making plus-sided errors only (or minus-sided errors only, resp.) is being closed under intersection (or closed under union, resp.).

Lemma 3. *Let $\mathcal{C} \subseteq \mathcal{H}$. If $\mathcal{H}$ contains hypotheses with plus-sided errors and hypotheses with minus-sided errors w.r.t. $\mathcal{C}$, then $\theta(\mathcal{C},\mathcal{H}) \leq 2$.*

Proof. Consider a fixed but arbitrary choice of $\mathbb{P}, X', h^*$. Let h_{min} be the hypothesis in $V_{\mathcal{H}}(X',h^*)$ that errs on positive examples of h^* only, and let h_{max} be the hypothesis in $V_{\mathcal{H}}(X',h^*)$ that errs on negative examples of h^* only. We conclude that $\mathrm{DIS}(V_{\mathcal{C}}(X',h^*)) \subseteq \{x |\ h_{min}(x) \neq h_{max}(x)\}$. From this and the triangle inequality, it follows that

$$\mathbb{P}(\mathrm{DIS}(V_{\mathcal{C}}(X',h^*))) \leq \mathbb{P}(h_{min} \neq h_{max}) \leq \mathbb{P}(h_{min} \neq h^*) + \mathbb{P}(h_{max} \neq h^*) \ .$$

The claim made by the lemma is now obvious from the definition of $\theta(\mathcal{C},\mathcal{H})$. □

Example 1. Since POWERSET and HALFINTERVALS are closed under intersection and union, we obtain $\theta(\mathrm{POWERSET}) \leq 2$ and $\theta(\mathrm{HALFINTERVALS}) \leq 2$. Let the class $\mathcal{C}$ consist of both the open and the closed homogeneous halfplanes and let $\mathcal{H}$ be the class of unions and intersections of two halfplanes from $\mathcal{C}$. It is easy to see that $\mathcal{H}$ contains hypotheses with plus-sided errors (the smallest pie slice with apex at $\mathbf{0}$ that includes all positive examples in a sample) and hypotheses with minus-sided errors (the complement of the smallest pie slice with apex at $\mathbf{0}$ that includes all negative examples in a sample) w.r.t. $\mathcal{C}$. Thus, $\theta(\mathcal{C},\mathcal{H}) \leq 2$. Note that $\mathcal{H}$ is neither closed under intersection nor closed under union.

3 A Closer Look to the Disagreement Coefficient

In Section 3.1 we investigate the question how small the product $\mathrm{VCdim}(\mathcal{C}) \cdot \theta(\mathcal{C}, \mathcal{H})$ can become if $\mathcal{H} \supseteq \mathcal{C}$ is cleverly chosen. The significance of this question should be clear from Lemma 2. In Section 3.2 we introduce a padding technique which leaves the disagreement coefficient invariant but increases the VC-dimension (and, as we will see later, also increases the error rates in the PAC Co-training Model).

3.1 A Combinatorial Upper Bound

Let $s^+(\mathcal{C})$ denote the largest number of instances in X such that every binary pattern on these instances with exactly one "+"-label can be realized by a concept from $\mathcal{C}$. In other words: $s^+(\mathcal{C})$ denotes the cardinality of the largest singleton-subclass of $\mathcal{C}$. Let $\mathcal{C}^+$ denote the class of all unions of concepts from $\mathcal{C}$. As usual, the empty union is defined to be the empty set.

Lemma 4. *$\mathcal{C} \subseteq \mathcal{C}^+$, $\mathcal{C}^+$ is closed under union, and $\mathrm{VCdim}(\mathcal{C}^+) = s^+(\mathcal{C})$. Moreover, if $\mathcal{C}$ is closed under intersection, then $\mathcal{C}^+$ is closed under intersection too, and $\theta(\mathcal{C}, \mathcal{C}^+) \le 2$ so that $\mathrm{VCdim}(\mathcal{C}^+) \cdot \theta(\mathcal{C}, \mathcal{C}^+) \le 2s^+(\mathcal{C})$.*

Proof. By construction, $\mathcal{C} \subseteq \mathcal{C}^+$ and $\mathcal{C}^+$ is closed under union. From this it follows that $s^+(\mathcal{C}) \le \mathrm{VCdim}(\mathcal{C}^+)$. Consider now instances $x_1, \ldots, x_d$ that are shattered by $\mathcal{C}^+$. Thus, for every $i = 1, \ldots, d$, there exists a concept h_i in $\mathcal{C}^+$ that contains x_i but none of the other $d-1$ instances. Therefore, by the construction of $\mathcal{C}^+$, $\mathcal{C}$ must contain some hypothesis h_i' smaller than h_i satisfying $h_i'(x_i) = 1$. We conclude that $\mathrm{VCdim}(\mathcal{C}^+) \le s^+(\mathcal{C})$. For the remainder of the proof, assume that $\mathcal{C}$ is closed under intersection. Consider two sets A, B of the form $A = \cup_i A_i$ and $B = \cup_j B_j$ where all A_i and B_j are concepts in $\mathcal{C}$. Then, according to the distributive law, $A \cap B = \cup_{i,j} A_i \cap B_j$. Since $\mathcal{C}$ is closed under intersection, $A_i \cap B_j \in \mathcal{C} \subseteq \mathcal{C}^+$. We conclude that $\mathcal{C}^+$ is closed under intersection. Closure under intersection and union implies that $\mathcal{C}^+$ contains hypotheses with plus-sided errors and hypotheses with minus-sided errors w.r.t. $\mathcal{C}$. According to Lemma 3, $\theta(\mathcal{C}, \mathcal{C}^+) \le 2$. □

We aim at a similar result that holds for arbitrary (not necessarily intersection-closed) concept classes. To this end, we proceed as follows. Let $s^-(\mathcal{C})$ denote the largest number of instances in X such that every binary pattern on these instances with exactly one "−"-label can be realized by a concept from $\mathcal{C}$. In other words: $s^-(\mathcal{C})$ denotes the cardinality of the largest co-singleton subclass of $\mathcal{C}$. Let $\mathcal{C}^-$ denote the class of all intersections of concepts from $\mathcal{C}$. As usual, the empty intersection is defined to be the full set X. By duality, Lemma 4 translates into the following

Corollary 1. *$\mathcal{C} \subseteq \mathcal{C}^-$, $\mathcal{C}^-$ is closed under intersection, and $\mathrm{VCdim}(\mathcal{C}^-) = s^-(\mathcal{C})$. Moreover, if $\mathcal{C}$ is closed under union, then $\mathcal{C}^-$ is closed under union too, and $\theta(\mathcal{C}, \mathcal{C}^-) \le 2$ so that $\mathrm{VCdim}(\mathcal{C}^-) \cdot \theta(\mathcal{C}, \mathcal{C}^-) \le 2s^-(\mathcal{C})$.*

We now arrive at the following general bound:

Theorem 1. *Let* $\mathcal{H} := \mathcal{C}^+ \cup \mathcal{C}^-$. *Then,* $\mathcal{C} \subseteq \mathcal{H}$, $\mathrm{VCdim}(\mathcal{H}) \le 2\max\{s^+(\mathcal{C}), s^-(\mathcal{C})\}$, *and* $\theta(\mathcal{C},\mathcal{H}) \le 2$ *so that* $\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C},\mathcal{H}) \le 4\max\{s^+(\mathcal{C}), s^-(\mathcal{C})\} =: s(\mathcal{C})$.

Proof. $\mathcal{C} \subseteq \mathcal{H}$ is obvious. The bound on the VC-dimension is obtained as follows. If m instances are given, then, by Lemma 4 and Corollary 1, the number of binary patterns imposed on them by concepts from $\mathcal{H} = \mathcal{C}^+ \cup \mathcal{C}^-$ is bounded by $\Phi_{s^+(\mathcal{C})}(m) + \Phi_{s^-(\mathcal{C})}(m)$ where

$$\Phi_d(m) = \begin{cases} 2^m & \text{if } m \le d \\ \sum_{i=0}^{d} \binom{m}{i} & \text{otherwise} \end{cases}$$

is the upper bound from Sauer's Lemma [10]. Note that $\Phi_d(m) < 2^{m-1}$ for $m > 2d$. Thus, for $m > 2\max\{s^+(\mathcal{C}), s^-(\mathcal{C})\}$, $\Phi_{s^+(\mathcal{C})}(m) + \Phi_{s^-(\mathcal{C})}(m) < 2^{m-1} + 2^{m-1} = 2^m$. We can conclude that $\mathrm{VCdim}(\mathcal{H}) \le 2\max\{s^+(\mathcal{C}), s^-(\mathcal{C})\}$. Finally note that $\theta(\mathcal{C},\mathcal{H}) \le 2$ follows from Lemma 3 and the fact that, because of Lemma 4 and Corollary 1, $\mathcal{H} = \mathcal{C}^+ \cup \mathcal{C}^-$ contains hypotheses with plus-sided errors and hypotheses with minus-sided errors. □

Please note that the parameter $s^-(\mathcal{C})$ was originally introduced by Mihály Geréb-Graus in [8] as the "unique negative dimension" of $\mathcal{C}$. He showed that it characterizes PAC-learnability from positive examples alone.

3.2 Invariance of the Disagreement Coefficient under Padding

For every domain X, let $X^{(i)}$ and $X^{[k]}$ be given by

$$X^{(i)} = \{(x,i) |\ x \in X\} \text{ and } X^{[k]} = X^{(1)} \cup \cdots \cup X^{(k)} \ .$$

For every concept $h \subseteq X$, let $h^{(i)} = \{(x,i) |\ x \in h\}$. For every concept class $\mathcal{C}$ over domain X, let

$$\mathcal{C}^{[k]} := \{h_1^{(1)} \cup \cdots \cup h_k^{(k)} |\ h_1, \ldots, h_k \in \mathcal{C}\} \ .$$

Loosely speaking, $\mathcal{C}^{[k]}$ contains k-fold "disjoint unions" of concepts from $\mathcal{C}$. It is obvious that $\mathrm{VCdim}(\mathcal{C}^{[k]}) = k \cdot \mathrm{VCdim}(\mathcal{C})$. The following result shows that the disagreement-coefficient is invariant under k-fold disjoint union:

Lemma 5. *For all* $k \ge 1$: $\theta(\mathcal{C}^{[k]}, \mathcal{H}^{[k]}) = \theta(\mathcal{C},\mathcal{H})$.

Proof. The probability measures $\mathbb{P}$ on $X^{[k]}$ can be written as convex combinations of probability measures on the $X^{(i)}$, i.e., $\mathbb{P} = \lambda_1 \mathbb{P}_1 + \cdots + \lambda_k \mathbb{P}_k$ where $\mathbb{P}_i$ is a probability measure on $X^{(i)}$, and the λ_i are non-negative numbers that sum-up to 1. A sample $S \subseteq X^{[k]}$ decomposes into $S = S^{(1)} \cup \cdots \cup S^{(k)}$ with

$S^{(i)} \subseteq X^{(i)}$. An analogous remark applies to concepts $c \in \mathcal{C}^{[k]}$ and hypotheses $h \in \mathcal{H}^{[k]}$. Thus,

$$\begin{aligned} \theta(\mathcal{C}^{[k]}, \mathcal{H}^{[k]}|\mathbb{P}, S, c) &= \frac{\mathbb{P}(\mathrm{DIS}(V_{\mathcal{C}^{[k]}}(S,c)))}{\sup_{h \in V_{\mathcal{H}^{[k]}}(S,c)} \mathbb{P}(h \neq c)} \\ &= \frac{\sum_{i=1}^k \lambda_i \overbrace{\mathbb{P}_i(\mathrm{DIS}(V_{\mathcal{C}^{(i)}}(S^{(i)}, c^{(i)})))}^{=:a_i}}{\sum_{i=1}^k \lambda_i \underbrace{\sup_{h_i^{(i)} \in V_{\mathcal{H}^{(i)}}(S^{(i)}, c^{(i)})} \mathbb{P}_i(h^{(i)} \neq c^{(i)})}_{=:b_i}} \leq \theta(\mathcal{C}, \mathcal{H}) \enspace . \end{aligned}$$

The last inequality holds because, obviously, $a_i/b_i \leq \theta(\mathcal{C}^{(i)}, \mathcal{H}^{(i)}) = \theta(\mathcal{C}, \mathcal{H})$. On the other hand, a_i/b_i can be made equal (or arbitrarily close) to $\theta(\mathcal{C}, \mathcal{H})$ by choosing $\mathbb{P}_i, S^{(i)}, c^{(i)}$ properly. □

4 Supervised Learning and Co-training

Let $p_+ = \mathbb{P}(h^* = 1)$ denote the probability for seeing a positive example of h^*. Similarly, $p_- = \mathbb{P}(h^* = 0)$ denotes the probability for seeing a negative example of h^*. Let $\mathbb{P}(\cdot|+), \mathbb{P}(\cdot|-)$ denote probabilities conditioned to positive or to negative examples, respectively. The error probability of a hypothesis h decomposes into conditional error probabilities according to

$$\mathbb{P}(h \neq h^*) = p_+ \cdot \mathbb{P}(h \neq h^*|+) + p_- \cdot \mathbb{P}(h \neq h^*|-) \enspace . \tag{2}$$

In the PAC-learning framework, a sample size that, with high probability, bounds the error by ε typically bounds the plus-conditional error by ε/p_+ and the minus-conditional error by ε/p_-. According to (2), these conditional error terms lead to an overall error that is bounded by ε, indeed. For this reason, the hardness of a problem in the PAC-learning framework does not significantly depend on the values of p_+, p_-. As we will see shortly, the situation is much different in the PAC Co-training Model under the Conditional Independence Assumption where small values of $p_{min} := \min\{p_+, p_-\}$ (though not smaller than ε) make the learning problem harder. Therefore, we refine the analysis and present our bounds on the sample size not only in terms of distribution-independent quantities like θ, ε and the VC-dimension but also in terms of p_{min}. This will lead to "smart" learning policies that take advantage of "benign values" of p_{min}. In the following subsections, we present (almost tight) upper and lower bounds on the sample size in the PAC Co-training Model under the Conditional Independence Assumption.

4.1 General Upper Bounds on the Sample size

Let us first fix some more notation that is also used in subsequent sections. $V_1 \subseteq \mathcal{C}_1$ and $V_2 \subseteq \mathcal{C}_2$ denote the version spaces induced by the labeled sample within the concept classes, respectively, and $\mathrm{DIS}_1 = \mathrm{DIS}(V_1)$, $\mathrm{DIS}_2 = \mathrm{DIS}(V_2)$ are the corresponding disagreement regions. The VC-dimension of $\mathcal{H}_1$ is denoted d_1; the VC-dimension of $\mathcal{H}_2$ is denoted d_2. $\theta_1 = \theta(\mathcal{C}_1, \mathcal{H}_1)$ and $\theta_2 = \theta(\mathcal{C}_2, \mathcal{H}_2)$. $\theta_{min} = \min\{\theta_1, \theta_2\}$ and $\theta_{max} = \max\{\theta_1, \theta_2\}$. $s_1^+ = s^+(\mathcal{C}_1)$, $s_2^+ = s^+(\mathcal{C}_2)$, $s_1^- = s^-(\mathcal{C}_1)$, and $s_2^- = s^-(\mathcal{C}_2)$. The learner's empirical estimates for p_+, p_-, p_{min} (inferred from the labeled random sample) are denoted $\hat{p}_+, \hat{p}_-, \hat{p}_{min}$, respectively. Let $h_1 \in V_{\mathcal{H}_1}$ and $h_2 \in V_{\mathcal{H}_2}$ denote two hypotheses chosen according to some arbitrary but fixed learning rules.

The error probability of the learner is the probability for erring on an unlabeled "test-instance" (x_1, x_2). Note that the learner has a safe decision if $x_1 \notin \mathrm{DIS}_1$ or $x_2 \notin \mathrm{DIS}_2$. As for the case $x_1 \in \mathrm{DIS}_1$ and $x_2 \in \mathrm{DIS}_2$, the situation for the learner is ambiguous, and we consider the following resolution-rules, the first two of which depend on the hypotheses h_1 and h_2: [5]

R1: If $h_1(x_1) = h_2(x_2)$, then vote for the same label. If $h_1(x_1) \neq h_2(x_2)$, then go with the hypothesis that belongs to the class with the disagreement coefficient θ_{max}.

R2: If $h_1(x_1) = h_2(x_2)$, then vote for the same label. If $h_1(x_1) \neq h_2(x_2)$, then vote for the label that occurred less often in the sample (i.e., vote for "+" if $\hat{p}_- \geq 1/2$, and for "−" otherwise).

R3: If $\hat{p}_- \geq 1/2$, then vote for label "+". Otherwise, vote for label "−". (These votes are regardless of the hypotheses h_1, h_2.)

Theorem 2. *The number of labeled examples sufficient for learning $(\mathcal{C}_1, \mathcal{C}_2)$ in the PAC Co-training Model under the Conditional Independence Assumption by learners applying one of the rules R1, R2, R3 is given asymptotically as follows:*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if rule R1 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \max\left\{\frac{1}{p_{min}}, \theta_{max}\right\}}\right) & \text{if rule R2 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1 \theta_2}\right) & \text{if rule R3 is applied} \end{cases} \tag{3}$$

Proof. $\tilde{O}(1)$ examples are sufficient to achieve that (with high probability) the following holds: if $p_{min} < 1/4$, then $\hat{p}_{min} < 1/2$. Assume that this is the case. For reasons of symmetry, we may assume furthermore that $\theta_1 = \theta_{max}$ and $\hat{p}_- \geq 1/2$ so that $p_- \geq 1/4$. Please recall that the rules R1 to R3 are only applied if $x_1 \in \mathrm{DIS}_1$ and $x_2 \in \mathrm{DIS}_2$. Assume first that ambiguities are resolved according

[5] The choice applied in rules R2 and R3 could seem counterintuitive at first. However, $\hat{p}_+ > \hat{p}_-$ means that the learner has more information about the behaviour of the target concept on the positive instances than on the negative ones, indicating that the positive instances in the disagreement regions might have smaller probability than the negative ones. This choice is also in accordance with the common strategy applied in the "learning from positive examples only" model, which outputs a negative label if in doubt, although the learner has never seen any negative examples.

to rule R1. Note that the sample size specified in (3) is sufficient to bound (with high probability) the error rate of hypotheses h_1, h_2, respectively, as follows [5]:

$$\varepsilon_1 = \sqrt{\frac{d_1}{d_2} \cdot \frac{p_{min}}{\theta_{min}} \cdot \varepsilon} \text{ and } \varepsilon_2 = \sqrt{\frac{d_2}{d_1} \cdot \frac{p_{min}}{\theta_{min}} \cdot \varepsilon}$$

If R1 assigns a wrong label to (x_1, x_2), then, necessarily, h_1 errs on x_1 and $x_2 \in \mathrm{DIS}_2$. Thus the error rate induced by R1 is bounded (with high probability) as follows:

$$\begin{aligned}
&\mathbb{P}(h_1(x_1) = 0 \wedge x_2 \in \mathrm{DIS}_2|+)p_+ + \mathbb{P}(h_1(x_1) = 1 \wedge x_2 \in \mathrm{DIS}_2|-)p_- \\
\leq\ & \frac{1}{p_{min}} \cdot \Big(\mathbb{P}(h_1(x_1) = 0|+)p_+ \cdot \mathbb{P}(x_2 \in \mathrm{DIS}_2|+)p_+ \\
&+ \mathbb{P}(h_1(x_1) = 1|-)p_- \cdot \mathbb{P}(x_2 \in \mathrm{DIS}_2|-)p_-\Big) \\
\leq\ & \frac{1}{p_{min}} \cdot \underbrace{\big(\mathbb{P}(h_1(x_1) = 0|+)p_+ + \mathbb{P}(h_1(x_1) = 1|-)p_-\big)}_{\leq \varepsilon_1} \\
&\cdot \underbrace{\big(\mathbb{P}(x_2 \in \mathrm{DIS}_2|+)p_+ + \mathbb{P}(x_2 \in \mathrm{DIS}_2|-)p_-\big)}_{\leq \theta_2 \varepsilon_2 = \theta_{min} \varepsilon_2} \leq \frac{\theta_{min}}{p_{min}} \cdot \varepsilon_1 \varepsilon_2 = \varepsilon
\end{aligned}$$

The first inequality in this calculation makes use of Conditional Independence and the third applies Lemma 2.
The proofs for the rules R2 and R3 proceed analogously. We omit the details because of space constraints. □

We now describe a strategy named "Combined Rule" that uses rules R1, R2, R3 as sub-routines. Given $(x_1, x_2) \in \mathrm{DIS}_1 \times \mathrm{DIS}_2$, it proceeds as follows. If $\varepsilon > 4/(\theta_1\theta_2)$ and $\hat{p}_+ \leq \varepsilon/2$ (or $\hat{p}_- \leq \varepsilon/2$, resp.), it votes for label "−" (or for label "+", resp.). If $\varepsilon \leq 4/(\theta_1\theta_2)$ or $\hat{p}_{min} := \min\{\hat{p}_+, \hat{p}_-\} > \varepsilon/2$, then it applies the rule

$$\begin{cases} \text{R1 if } \frac{\theta_{min}}{\theta_{max}} \leq \hat{p}_{min} \\ \text{R2 if } \frac{1}{\theta_1\theta_2} \leq \hat{p}_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \text{R3 if } \hat{p}_{min} < \frac{1}{\theta_1\theta_2} \end{cases} . \tag{4}$$

Corollary 2. *If the learner applies the Combined Rule, then*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if } \frac{\theta_{min}}{\theta_{max}} \leq p_{min} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right) & \text{if } \frac{1}{\theta_{max}} \leq p_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{1}{p_{min}}}\right) & \text{if } \frac{1}{\theta_1\theta_2} \leq p_{min} < \frac{1}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1\theta_2}\right) & \text{if } p_{min} < \frac{1}{\theta_1\theta_2} \end{cases} \tag{5}$$

labeled examples are sufficient for learning $\mathcal{C}_1, \mathcal{C}_2$ in the PAC Co-training Model under the Conditional Independence Assumption.

Proof. It is an easy application of multiplicative Chernov-bounds to show that (with high probability) $\hat{p}_{min}$ equals p_{min} up to a factor of 2 unless one of the following special cases occurs:

$$\varepsilon > \frac{4}{\theta_1\theta_2} \ , \ \hat{p}_{min} < \frac{\varepsilon}{2} \ \text{ and } \ p_{min} < \frac{\varepsilon}{2} \tag{6}$$

$$\varepsilon \le \frac{4}{\theta_1\theta_2} \ , \ \hat{p}_{min} < \frac{1}{\theta_1\theta_2} \ \text{ and } \ p_{min} < \frac{1}{\theta_1\theta_2} \tag{7}$$

In case of (6), the Combined Rule outputs the empirically more likely label, and the error rate is bounded by $p_{min} < \varepsilon/2$. In case of (7), rule R3 is applied which, according to Theorem 2, leads to the desired upper bound on the sample size. Thus, we may now safely assume that $\hat{p}_{min}$ equals p_{min} up to factor 2. If none of the rules R1, R2, R3 is applied, then $\hat{p}_{min} \le \varepsilon/2$ and the error rate will be $p_{min} < \varepsilon$. Thus, we may assume that Combined Rule proceeds according to (4). If the learner could substitute the (unknown) p_{min} for $\hat{p}_{min}$ within (4), the corollary would follow immediately from Theorem 2. But it is easy to see that even the knowledge of the empirical estimate $\hat{p}_{min}$ is sufficient for this purpose. □

We can also give a completely combinatorial upper bound without referring to θ. As we will see later this bound is tight up to logarithmic factors in the worst case (i.e. small p_{min}):

Theorem 3. *If the learner applies rule R3 and uses hypothesis classes* $\mathcal{H}_b = \mathcal{C}_b^+ \cup \mathcal{C}_b^-$ *for* $b = 1, 2$*, then* $\tilde{O}\big(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}\big)$ *labeled examples are sufficient.*

Proof. (Sketch) R3 always chooses one of the two hypotheses with one sided-error: h^- which always outputs "–" for $(x_1, x_2) \in DIS_1 \times DIS_2$, and h^+ which always outputs "+" on these instances. With an analysis similar as before one can relate the error of h^- to the errors of the smallest consistent hypotheses in $\mathcal{C}_1^-$ and $\mathcal{C}_2^-$, and then conclude using standard PAC-bounds and results from Section 3.1 that with high probability $\tilde{O}\big(\sqrt{s_1^- s_2^-/(\varepsilon p_+)}\big)$ many examples suffice for h^- to have an error of at most ε. For h^+ the analogous bound on the number of examples is $\tilde{O}\big(\sqrt{s_1^+ s_2^+/(\varepsilon p_-)}\big)$. Because $\tilde{O}(1)$ many examples are enough to distinguish with high probability among the cases "p_- is large" (say $\ge 3/4$) and "p_- is small" (say $\le 1/4$), choosing between the two learning strategies according to rule R3 yields the desired bound. □

Please note that finding the smallest consistent hypotheses in $\mathcal{C}_b^-$ is possible using positive examples only. This shows a strong connection to the results by Geréb-Graus in [8], where it was shown that $\tilde{O}\big(s^-(\mathcal{C})/\varepsilon\big)$ many positive examples are sufficient (and necessary) to PAC-learn a class $\mathcal{C}$ from positive examples alone. A dual also holds for the largest consistent hypotheses in C_b^+.

4.2 Lower Bounds on the Sample Size

In this section, we first derive a general lower bound which matches the upper bound from Theorem 3. Both bounds are part of a worstcase analysis and the lower bound makes use of rather small values of p_{min}. Afterwards, we show that, for more "benign" choices of p_{min}, the upper bound from Corollary 2 is tight by exhibiting concrete concept classes that lead to a matching lower bound.
A useful concept class for the purposes of this section is SF_n from Lemma 1. Note that all lower bounds obtained for SF_n immediately generalize to concept classes containing SF_n as subclass.

Lemma 6. *Let $n_1, n_2 \geq 1$, and let $\mathcal{C}_b = \mathrm{SF}_{n_b+2}$ so that $\theta_b = n_b + 2$ for $b = 1, 2$. Then, for every $p_{min} \leq 1/(\theta_1\theta_2)$ and every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1, \mathcal{C}_2$ in the PAC Co-training Model under the Conditional Independence Assumption is at least $\Omega(\sqrt{n_1 n_2/\varepsilon})$.*

Proof. Let $p_+ = p_{min}$, let $\mathcal{C}_1$ be a concept class over domain $\{a_0, a_1, \ldots, a_{n_1+2}\}$, and let $\mathcal{C}_1$ be a concept class over domain $\{b_0, b_1, \ldots, b_{n_2+2}\}$, respectively. Obviously,

$$p_+ = p_{min} \leq \frac{1}{(n_1+2)(n_2+2)} = \frac{1}{\theta_1\theta_2} \enspace .$$

Note that $n_1 n_2 p_+ \leq 1$. Consider the following malign scenario:

- $\mathbb{P}(a_0|+) = \mathbb{P}(b_0|+) = 1 - \sqrt{\varepsilon/p_+}$.
- Index s is uniformly chosen at random from $\{2, \ldots, n_1 + 2\}$. Index t is uniformly chosen at random from $\{2, \ldots, n_2+2\}$. $\mathbb{P}(a_s|+) = \mathbb{P}(b_t|+) = \sqrt{\varepsilon/p_+}$.
- $\mathbb{P}(a_1|-) = 1 - 4 \cdot \sqrt{n_1\varepsilon/n_2}$. $\mathbb{P}(b_1|-) = 1 - 4 \cdot \sqrt{n_2\varepsilon/n_1}$.
- The instances from $X_1 \setminus \{a_0, a_1, a_s\}$ evenly share a minus-conditional probability mass of $4 \cdot \sqrt{n_1\varepsilon/n_2}$. The instances from $X_2 \setminus \{b_0, b_1, b_t\}$ evenly share a minus-conditional probability mass of $4 \cdot \sqrt{n_2\varepsilon/n_1}$.

Let us assume that the sample size satisfies $m \leq \frac{\sqrt{n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon}$. Let Z_1 count the number of sample points that hit $X_1 \setminus \{a_0, a_1, a_s\}$ (the "interesting negative examples" in X_1). Let Z_2 be defined analogously. Then the following holds:

- The expectation of Z_b is bounded by $n_b/10$ for $b = 1, 2$. Thus, with probability at least $1 - 2/5$, $Z_1 \leq n_1/2$ and $Z_2 \leq n_2/2$, which is assumed in the sequel. Thus at least half of the interesting negative examples in X_1 and at least half of the interesting negative examples in X_2 remain "hidden" from the learner (i.e. do not occur in the sample), respectively.
- The expected number of occurrences of a_s (or b_t, resp.) in the sample is bounded by

  $$p_+ \cdot \sqrt{\varepsilon/p_+} \cdot \frac{\sqrt{n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon} = \sqrt{n_1 n_2 p_+}/40 \leq 1/40 \enspace .$$

  Thus, with probability at least $1 - 1/15$, neither a_s nor b_t occurs in the sample, which is assumed in the sequel.

Note that the assumptions that we made on the way are satisfied with a probability of at least $1 - 2/5 - 1/15 > 1/2$. Given these assumptions, we now bound the smallest possible error rate from below. Let E_+ (or E_-, resp.) denote the set of instance-pairs which are labeled "+" (or labeled "−", resp.). For $b = 1, 2$, let $U_b \subseteq X_b$ be the set of points in X_b that did not occur in the sample, and let $U = U_1 \times U_2$. For test-instances $(x_1, x_2) \notin U$, the learner can infer the label from the information provided by the sample. It can be shown by a rigorous analysis (omitted here because of space constraints) that the Bayes-decision leads to the same vote for all pairs from U: if $\mathbb{P}(E_+ \cap U) \geq \mathbb{P}(E_- \cap U)$ vote for "+", otherwise vote for "−". Clearly, the resulting Bayes-error equals $\min\{\mathbb{P}(E_+ \cap U), \mathbb{P}(E_- \cap U)\}$. It can be bounded from below as follows:

$$\mathbb{P}(U \cap E_+) \geq p_+ \cdot \left(\sqrt{\frac{\varepsilon}{p_+}}\right)^2 = \varepsilon \ ,$$

because $\sqrt{\frac{\varepsilon}{p_+}}$ coincides with the plus-conditional probability of a_s and b_t, respectively. A similar computation shows that

$$\mathbb{P}(U \cap E_-) \geq \underbrace{(1 - p_+)}_{\geq 1/2} \cdot (2\sqrt{n_1\varepsilon/n_2}) \cdot (2\sqrt{n_2\varepsilon/n_1}) \geq 2\varepsilon \ .$$

Thus, the Bayes-error is at least ε. □

Corollary 3. *Let $n_1, n_2, d_1, d_2 \geq 1$, and, for $b = 1, 2$, let $\mathcal{C}_b = \mathrm{SF}_{n_b+2}$ so that $\theta(\mathcal{C}_b) = \theta(\mathcal{C}_b^{[d_b]}) = n_b + 2$. Then, for every $p_{min} \leq 1/(\theta_1\theta_2)$ and every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1^{[d_1]}, \mathcal{C}_2^{[d_2]}$ in the PAC Co-training Model under the Conditional Independence Assumption is at least $\Omega(\sqrt{d_1 d_2 n_1 n_2/\varepsilon})$.*

Proof. (Sketch) $\theta(\mathcal{C}_b) = \theta(\mathcal{C}_b^{[d_b]})$ follows from Lemma 5. A malign scenario for the classes $\mathcal{C}_b^{[d_b]}$ is obtained by installing the malign scenario from the proof of Lemma 6 (with some minor modifications) for each of the d_1 many copies of $\mathcal{C}_1$ and for each of the d_2 many copies of $\mathcal{C}_2$. The main idea behind the proof is that every disjoint copy of the "old scenario" is now served by fewer sample points. In order to compensate this, the sample size must pop-up by factor $\sqrt{d_1 d_2}$. □

Here comes the lower bound that is tight from the perspective of a worstcase analysis:

Theorem 4. *Assume that*[6] $3 \leq s_b^+, s_b^- < \infty$ *for $b \in \{0, 1\}$. Then the following holds. For every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1, \mathcal{C}_2$ in the PAC Co-training Model under the Conditional Independence Assumption is at least $\Omega\left(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}\right)$.*

[6] One can drop the restriction $3 \leq s_b^+, s_b^-$ and still prove tight bounds, but that needs a tedious case distinction and is omitted due to space constraints.

Proof. We show first that each learner needs at least $\Omega\big(\sqrt{s_1^+ s_2^+/\varepsilon}\big)$ many labels in the worst case. By duality we also get the bound of $\Omega\big(\sqrt{s_1^- s_2^-/\varepsilon}\big)$. Now taking the maximum of both cases yields the theorem.

The former bound can be proved as follows: in the proof of Lemma 6, we may set $p_+ = p_{min} = \varepsilon$. Thus the probability assigned to the redundant points a_0 and b_0 is now 0, respectively. Removal of the redundant point in a class of type SF will lead to the class of singletons. Thus, the proof of Lemma 6 with the special setting $p_+ = p_{min} = \varepsilon$ shows that at least $\Omega\big(\sqrt{n_1 n_2/\varepsilon}\big)$ many examples are needed for every pair $\mathcal{C}_1, \mathcal{C}_2$ of concept classes such that, for $b = 1, 2$, $\mathcal{C}_b$ contains a singleton subclass of size $n_b + 2$. □

Note that the lower bound in Theorem 4 nicely matches with the upper bound in Theorem 3 for small enough ε. Furthermore, this implies a weak converse of Theorem 1 where we have shown that $\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C}, \mathcal{H}) \leq s(\mathcal{C})$ for $\mathcal{H} = \mathcal{C}^+ \cup \mathcal{C}^-$. More precisely $s(\mathcal{C}) = \tilde{O}\big(\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C}, \mathcal{H})\big)$ must hold for every $\mathcal{H} \supseteq \mathcal{C}$ because, otherwise, the lower bound in Theorem 4 would exceed the upper bound for rule R3 in Theorem 2 with $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$.

The next step will be to provide lower bounds that remain valid even when p_{min} takes more "benign values" than it does in the worstcase. Actually, Corollary 3 is a first step in this direction because the lower bound in this result nicely matches with the upper bound in Corollary 2 when $p_{min} \leq 1/(\theta_1\theta_2)$. We list here, without proof, some more lemmas of this kind which together witness that all upper bounds mentioned in Corollary 2 are fairly tight.

For any concept class $\mathcal{C}$ over domain X, the class $\mathrm{co}(\mathcal{C})$ is given by $\mathrm{co}(\mathcal{C}) = \{X \setminus A |\ A \in \mathcal{C}\}$. Clearly, $\mathrm{VCdim}(\mathcal{C}) = \mathrm{VCdim}(\mathrm{co}(\mathcal{C}))$ and $\theta(\mathcal{C}) = \theta(\mathrm{co}(\mathcal{C}))$.

Lemma 7. *Let $k, n \geq 1$, let $\mathcal{C}_1 = \mathrm{SF}_{kn+2}$, and let $\mathcal{C}_2 = \mathrm{co}(\mathrm{SF}_{n+2})$, so that $\theta_{max} = \theta(\mathcal{C}_1) = \theta(\mathcal{C}_1^{[d_1]}) = kn+2$ and $\theta_{min} = \theta(\mathcal{C}_2) = \theta(\mathcal{C}_2^{[d_2]}) = n+2$. Then, for every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1^{[d_1]}, \mathcal{C}_2^{[d_2]}$ in the PAC Co-training Model under the Conditional Independence Assumption is at least*

$$\begin{cases} \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if } \frac{\theta_{min}}{\theta_{max}} \leq p_{min} \leq \frac{1}{2} \\ \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right) & \text{if } \frac{1}{\theta_{max}} \leq p_{min} \leq \frac{\theta_{min}}{\theta_{max}} \end{cases} .$$

Lemma 8. *Let $n \geq 2$, let $\mathcal{C}_1 = \mathcal{C}_2 = \mathrm{SF}_{n+2}$, so that $\theta_1 = \theta_2 = \theta_{max} = n+2$. Then, for every $1/(\theta_1\theta_2) \leq p_{min} \leq 1/\theta_{max}$ and every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1^{[d_1]}, \mathcal{C}_2^{[d_2]}$ in the PAC Co-training Model under the Conditional Independence Assumption is at least $\Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon p_{min}}}\right)$.*

The lower bounds in Corollary 3 and Lemmas 7 and 8 nicely match with the general upper bounds given in Corollary 2.

4.3 Sample Size in Case of One-sided Errors

In the upper bounds presented in this section, any term of the form $d_b\theta_b$ can be safely replaced by $s(\mathcal{C}_b)$ provided that $\mathcal{H}_b = \mathcal{C}_b^+ \cup \mathcal{C}_b^-$. The proofs will be presented in the journal version.

Theorem 5. *For $b = 1, 2$, let $\mathcal{C}_b, \mathcal{H}_b$ be classes such that $\mathcal{H}_b$ contains hypotheses with plus-sided errors (or with minus-sided errors, resp.) w.r.t. $\mathcal{C}_b$. Then sample size*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{1}{p_{min}}}\right) & \text{if } p_{min} \geq \frac{1}{\theta_1\theta_2} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1\theta_2}\right) & \text{otherwise} \end{cases}$$

is asymptotically sufficient for learning $\mathcal{C}_1, \mathcal{C}_2$ with hypotheses from $\mathcal{H}_1, \mathcal{H}_2$ in the PAC Co-training Model under the Conditional Independence Assumption.

Note that the upper bound from Theorem 5 applies to the special case where, for $b = 1, 2$, $\mathcal{H}_b = \mathcal{C}_b$ and $\mathcal{C}_b$ is intersection-closed (or union-closed, resp.). In this case, the upper bound nicely matches with the lower bounds from Corollary 3 and Lemma 8.

Theorem 6. *For $b = 1, 2$, let $\mathcal{C}_b, \mathcal{H}_b$ be classes such that $\mathcal{H}_1$ contains hypotheses with plus-sided errors w.r.t. $\mathcal{C}_1$, and $\mathcal{H}_2$ contains hypotheses with minus-sided errors w.r.t. $\mathcal{C}_2$. Then sample size*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if } p_{min} \geq \frac{\theta_{min}}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right) & \text{otherwise} \end{cases}$$

is asymptotically sufficient for learning $\mathcal{C}_1, \mathcal{C}_2$ with hypotheses from $\mathcal{H}_1, \mathcal{H}_2$ in the PAC Co-training Model under the Conditional Independence Assumption.

Note that the upper bound from Theorem 6 applies to the special case where $\mathcal{H}_1 = \mathcal{C}_1$ is intersection-closed and $\mathcal{H}_2 = \mathcal{C}_2$ is union-closed. In this case, the upper bound nicely matches with the lower bound from Lemma 7.

5 Final Remarks

It is known that semi-supervised learners in the Co-training framework also benefit from assumptions weaker than conditional independence (see [2, 12]). One can ask whether PAC-learners can also use these relaxed assumptions to their advantage. At least for the assumption introduced in [2] this is not generally true: for Co-training with an α-expanding distribution and one-sided errors, one can show that there are classes and distributions (e.g. "Example 1" from [2]) where every PAC-learner requires $\Omega(d/\varepsilon)$ many examples (with d denoting the VC-dimension of both views), which coincides with the standard PAC bounds. On the other hand, conditional independence given the label reduces the label complexity to $\tilde{O}(d/\sqrt{\varepsilon})$. Details will follow in the journal version of this paper.

We also looked into having more than two views. With k views under the Conditional Independence Assumption we can show that the upper bound for rule R3 becomes $m = \tilde{O}\big(\sqrt[k]{d_1\theta_1 \cdots d_k\theta_k/\varepsilon}\big)$, and, as in the 2-view case, this has a matching lower bound. The other bounds can be generalized in a similar fashion.

The lower bound given in Theorem 4 is only valid for finite s_b^+, s_b^-, because the constraint on ε is essentially $1/\varepsilon \geq \max\{s_1^+ s_2^+, s_1^- s_2^-\}$. In case the singleton size is infinite, however, this theorem still implies the lower bound $\Omega(1/\varepsilon)$. Nevertheless, this rules out the drastic reduction of the label complexity that we saw for $s_b^+, s_b^- < \infty$. To determine how much the Co-training assumption can help in this situation is work in progress.

In a broader context, it would be interesting to see whether the techniques of this paper can be applied to get new bounds on the unlabeled sample complexity in semi-supervised learning. Another interesting question is whether existing upper bounds in active learning (at least in the realizable case) can be reformulated in completely combinatorial terms using Theorem 1.

References

[1] 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.

[2] Marina-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems 17*, pages 89–96. MIT Press, 2005.

[3] 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.

[4] 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.

[5] 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.

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

[7] Malte Darnstädt and Hans U. Simon. Smart PAC-learners. *Theoretical Computer Science*, 412(19):1756–1766, 2011.

[8] Mihály Geréb-Graus. Lower bounds on parallel, distributed and automata computations. PhD thesis, Harvard University Cambridge, MA, USA, 1989.

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

[10] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

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

[12] Wei Wang and Zhi-Hua Zhou. A new analysis of co-training. In *ICML*, pages 1135–1142, 2010.