Supervised Learning and Co-training[☆]

Malte Darnstädt[a], Hans Ulrich Simon[a,*], Balázs Szörényi[b,c]

[a] *Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany*
[b] *Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence, H-6720 Szeged, Hungary*
[c] *INRIA Lille, SequeL project, F-59650 Villeneuve d'Ascq, France*

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 in the realizable case 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.

Keywords: PAC-learning, Co-training, supervised learning

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.[1] 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, 9].

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 [5].

1.1. Related work

Supervised and semi-supervised learning. In the semi supervised learning framework the learner is assumed to have access both to labeled and unlabeled data. The former is supposed to be expensive and the latter to be cheap, thus unlabeled data should be used to minimize the amount of labeled data required. Indeed, a large set of unlabeled data provides extra information about the underlying distribution.

Already in 1991, Benedek and Itai [4] studied learning under a fixed distribution, which can be seen as an extreme case of semi-supervised learning, where the learner has full knowledge of the underlying

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

[*] Corresponding author

[1] See the introduction of [8] for a discussion of the most popular assumptions.

distribution. They derive upper and lower bounds on the number of required labels based on ε-covers and -packings. Later in 2005, Kääriäinen [13] developed a semi-supervised learning strategy, which can save up to one half of the required labels. These results don't make use of extra assumptions that relate the target concept to the data distribution.

However, some recent results by Ben-David et al. in [3] and later by Darnstädt and Simon in [9] indicate that even knowing the data distribution perfectly does not help the learner for most distributions asymptotically, i.e. a reduction by a constant factor is the best possible. In fact, they conjecture a general negative result, which is nonetheless still absent. These results can be regarded as a justification of using extra assumptions in the semi-supervised framework in order to make real use of having access to unlabeled data.

Our work provides a similar analysis of these assumptions in the fashion of the above results: we investigate to what extent does such an assumption (Co-training with the Conditional Independence Assumption) alone help the learner, and how much is to be credited to having perfect knowledge about the underlying distribution.

Likewise, a study for the popular Cluster Assumption was done by Singh, Nowak and Zhu in [15]. They show that the value of unlabeled data under their formalized Cluster Assumption varies with the minimal margin between clusters.

Co-training and the Conditional Independence Assumption. The co-training model was introduced by Blum and Mitchell in [5], and has an extensive literature in the semi-supervised setting, especially from an empirical and practical point of view. (For the formal definition see Section 2.) A theoretical analysis of Co-training under the Conditional Independence Assumption [5], and the weaker α-expanding Assumption [2], was accomplished by Balcan and Blum in [1]. They work in Valiant's model of PAC-learning [16] and show that *one labeled example* is enough for achieving the success criterion of PAC-learning provided that there are sufficiently many unlabeled examples.[2]

Our paper complements their results: we also work in the PAC model and prove label complexity bounds, but in our case the learner has no access to unlabeled data. As far as we know, our work is the first that studies Co-training in a fully supervised setting. Assuming Conditional Independence, our label complexity bound is much smaller than the standard PAC bound (which must be solely awarded to Co-training itself), while it is still larger than Balcan and Blum's (which must also be awarded to the use of unlabeled data). See Section 1.2 for more details.

Agnostic active learning. We make extensive use of a suitably defined variant of Hanneke's disagreement coefficient, which was introduced in [11] to analyze agnostic active learning. (See Section 2.2 for a comparison of the two notions.) To our knowledge this is, besides a remark about classical PAC-learning in Hanneke's thesis [12], the first use of the disagreement coefficient outside of agnostic learning. Furthermore, our work doesn't depend on results from the active learning community, which makes the prominent appearance of the disagreement coefficient even more remarkable.

Learning from positive examples only. Another unsuspected connection that emerged from our analysis relates our work to the "learning from positive examples only" model from [10]. As already mentioned, we can upper bound the product of the VC-dimension and the disagreement coefficient by a combinatorial parameter that is strongly connected to Geréb-Graus' "unique negative dimension". Furthermore, we derive worst case lower bounds that make use of this parameter.

1.2. Our main result

Our paper is a continuation of the line of research started out by Ben-David et al. in [3] aiming at investigating the problem: how much can the learner benefit from knowing the underlying distribution. We

[2]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, 17] for suggestions of relaxed assumptions.

investigate this problem focusing on a popular assumption in the semi supervised literature. Our results are purely theoretical, which also stems from the nature of the problem.

As mentioned above, the model of Co-training under the Conditional Independence Assumption was introduced in [5] as a setting where semi supervised can be superior to fully supervised learning. Indeed, in [1] it was shown that a single labeled example suffices for PAC-learning in the realizable case if unlabeled data is available. Recall that supervised, realizable 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 [6]. 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.

1.3. Organization of the paper

The remainder of the paper is structured as follows. Section 2 gives a short introduction to PAC-learning, 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 [11] to a purely combinatorial parameter, $s(\mathcal{C})$, which is closely related to the "unique negative dimension" from [10]. 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 worst case analysis has one-sided error. Section 4.3 presents improved bounds for classes with special properties. Section 4.4 shows a negative result in the more relaxed model of Co-training with α-expansion. The closing Section 5 contains some final remarks and open questions.

2. Definitions, Notations, and Facts

We first want to recall a result from probability theory that we will use several times:

Theorem 1 (Chernoff-bounds). *Let $X_1, \ldots, X_m$ be a sequence of independent Bernoulli random variables, each with the same probability of success $p = \mathbb{P}(X_1 = 1)$. Let $S = X_1 + \ldots + X_m$ denote their sum. Then for $0 \leq \gamma \leq 1$ the following holds:*

$$\begin{aligned} \mathbb{P}(S > (1+\gamma) \cdot p \cdot m) &\leq e^{-mp\gamma^2/3} \\ \textit{and} & \\ \mathbb{P}(S < (1-\gamma) \cdot p \cdot m) &\leq e^{-mp\gamma^2/2} \end{aligned}$$

2.1. The PAC-learning framework

We will give a short introduction to Valiant's model of Probably Approximately Correct Learning (PAC-learning) [16] with results from [6]. This section can be skipped if the reader is already familiar with this model.

Let X be any set, called the domain, and $\mathbb{P}$ be a probability distribution over X. For any $m \geq 1$, $\mathbb{P}^m$ denotes the corresponding product measure. Let 2^X denote the power-set of X (set of all subsets of X). In learning theory a family $\mathcal{C} \subseteq 2^X$ of subsets of X is called a *concept class over domain* X. Members $c \in \mathcal{C}$ are sometimes viewed as functions from X to $\{0,1\}$ (with the obvious one-to-one correspondence between these functions and subsets of X). We call any family $\mathcal{H} \subseteq 2^X$ with $\mathcal{C} \subseteq \mathcal{H}$ a *hypothesis class for* $\mathcal{C}$. An algorithm $\mathcal{A}$ is said to *PAC-learn* $\mathcal{C}$ *by* $\mathcal{H}$ *with sample size* $m(\varepsilon, \delta)$ if, for any $0 < \varepsilon, \delta < 1$, any (so-called) target concept $h^* \in \mathcal{C}$, and any domain distribution $\mathbb{P}$ the following holds:

1. If $\mathcal{A}$ is applied to a (so-called) sample $(x_1, h^*(x_1)), \ldots, (x_m, h^*(x_m))$, it returns (a "natural" representation of) a hypothesis $h \in \mathcal{H}$.
2. If $m = m(\varepsilon, \delta)$ and the instances $x_1, \ldots, x_m$ in the sample are drawn at random according to $\mathbb{P}^m$, then, with probability at least $1 - \delta$, $\mathbb{P}(h(x) = 0 \wedge h^*(x) = 1) + \mathbb{P}(h(x) = 1 \wedge h^*(x) = 0) \leq \varepsilon$.

Obviously, the term $\mathbb{P}(h(x) = 0 \wedge h^*(x) = 1) + \mathbb{P}(h(x) = 1 \wedge h^*(x) = 0)$ is the probability of $\mathcal{A}$'s hypothesis h to err on an random example. We often call this probability the *error rate* of the learner. The number m of examples that the best algorithm needs to PAC-learn $\mathcal{C}$ by $\mathcal{H}$ for the worst case choice of $\mathbb{P}$ and h^* is called the *sample complexity of* $(\mathcal{C}, \mathcal{H})$. We briefly note that in the original definition of PAC-learning [16] $m(\varepsilon, \delta)$ is required to be polynomially bounded in $1/\varepsilon, 1/\delta$ and $\mathcal{A}$ has to be polynomially time-bounded. In this paper, we do obtain polynomial bounds on the sample size, but we do not care about computational or efficiency issues.

We will now state two classic results from the PAC-learning framework, which show that there are (up to logarithmic factors) matching upper and lower bounds on the sample complexity. To this end we need to introduce some more definitions:

We say that a set $A \subseteq X$ is *shattered by* $\mathcal{H}$ if, for any $B \subseteq A$, there exists a set $C \in \mathcal{H}$ such that $B = A \cap C$. The *VC-dimension* of $\mathcal{H}$ is ∞ if there exist arbitrarily large sets that are shattered by $\mathcal{H}$, and the cardinality of the largest set shattered by $\mathcal{H}$ otherwise.

For every $h^* \in \mathcal{C}$ and every $X' \subseteq X$, the corresponding *version space in* $\mathcal{H}$ is given by

$$V_{\mathcal{H}}(X', h^*) := \{h \in \mathcal{H} |\ \forall x \in X' : h(x) = h^*(x)\}\ .$$

For $X' = \{x_1, \ldots, x_m\}$, we call the hypotheses in $V_{\mathcal{H}}(X', h^*)$ *consistent with the sample* $(x_1, h^*(x_1))$, $\ldots$, $(x_m, h^*(x_m))$.

Theorem 2 ([6]). *Let $0 < \varepsilon, \delta < 1$ and let d denote the VC-dimension of $\mathcal{H}$. An algorithm, that returns a consistent hypothesis, achieves an error rate of at most ε with probably $1 - \delta$ after receiving a sample of size*

$$m = O\left(\frac{1}{\varepsilon} \cdot \left(\ln \frac{1}{\delta} + d \cdot \ln \frac{1}{\varepsilon}\right)\right)$$

as its input.

Thus we have an upper bound of $\tilde{O}(d/\varepsilon)$ on the sample complexity of learning $\mathcal{C}$ by $\mathcal{H}$. $\tilde{O}$ is defined like Landau's O but also hides logarithmic factors.

Theorem 3 ([7]). *Let $d_{\mathcal{C}}$ denote the VC-dimension of $\mathcal{C}$. For small enough $\epsilon, \delta > 0$, any algorithm learning $\mathcal{C}$ (by any $\mathcal{H}$) for all choices of h^* and $\mathbb{P}$ must use at least*

$$m = \Omega\left(\frac{1}{\varepsilon} \cdot \left(\ln \frac{1}{\delta} + d_{\mathcal{C}}\right)\right)$$

many sample points.

So if we choose a hypothesis class with a VC-dimension in the same order as the VC-dimension of the concept class (e.g. $\mathcal{H} = \mathcal{C}$), we have a lower bound of $\Omega(d/\varepsilon)$ on the sample size, which matches the upper bound up to logarithmic factors.

The standard PAC-learning model described above is fully supervised, in the sense that the learner is provided with the correct label $h^*(x_i)$ for each sample point x_i. In *semi-supervised learning*, which we mention several times in this paper, the learner has access to a second (usually large) sample of unlabeled data and tries to use her increased knowledge about the domain distribution $\mathbb{P}$ to reduce the number of needed labels. For an analysis of semi-supervised learning in an augmented version of the PAC-learning model we refer to [1].

2.2. Co-training and the disagreement coefficient

In Co-training [5], 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 $\mathbb{P}$, 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 [5, 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. Let $V \subseteq \mathcal{C}$. The *disagreement region of* V is given by

$$\mathrm{DIS}(V) := \{x \in X |\ \exists h, h' \in V : h(x) \neq h'(x)\}\ .$$

We define the following variants of disagreement coefficients:

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

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.

To become more familiar with these definitions we present the following example:

Example 1. *Let X be $\mathbb{R}^2$ and let $\mathcal{C} = \mathcal{H}$ be the class of homogeneous half planes. The target concept h^* is illustrated in Figure 1 by the dashed line. The learner now receives a sample X', which consists of finitely many points (we have eight in our example), chosen randomly according to distribution P. Let us assume in this example, that P is the uniform distribution on $[-1, 1] \times [-1, 1]$. A positive or negative label, produced by h^*, is attached to each sample point, which we represent in Figure 1 by "+" and "−".*

$V_{\mathcal{C}}(X', h^*)$, which is not shown explicitly in the figure, consists of all homogeneous half planes covering all positively labeled points from X' while excluding the negative ones. One arbitrarily chosen hypothesis h from $V_{\mathcal{C}}(X', h^*)$, which the learner may pick as her answer, is shown in the left image. The error of the hypothesis h is the weight of the gray area, on which h and h^* differ, measured by P.

The disagreement region, which consists of all points where some *hypothesis from $V_{\mathcal{C}}(X', h^*)$ disagrees with the target h^*, is shown in the right-hand part of Figure 1.*

The disagreement coefficient $\theta(\mathcal{C}, \mathcal{H} | \mathbb{P}, X', h^*)$ is the ratio of the size of the disagreement region, measured again by $\mathbb{P}$, and the largest possible error of a half plane consistent with the sample. Please note that in this example, for any choice of h^*, $\mathbb{P}$ and X', the largest error can never exceed the whole disagreement region, and therefore $\theta(\mathcal{C}, \mathcal{H}) \geq 1$.

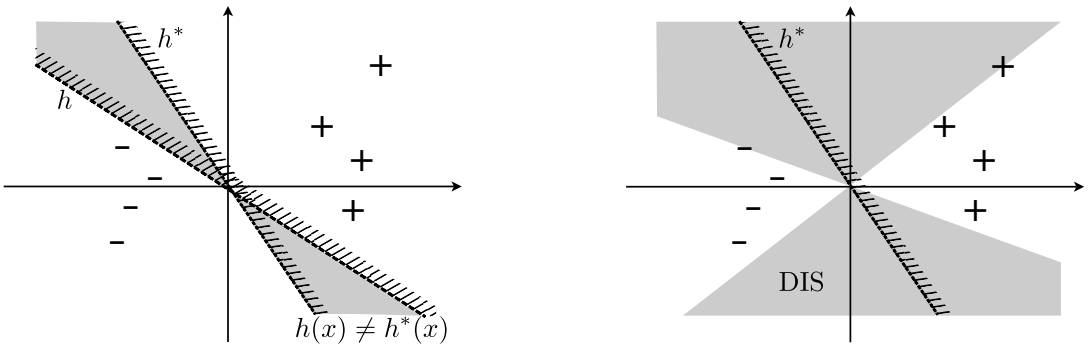

Figure 1: An illustration of h^*, X', h, the error of h and the disagreement region for Example 1.

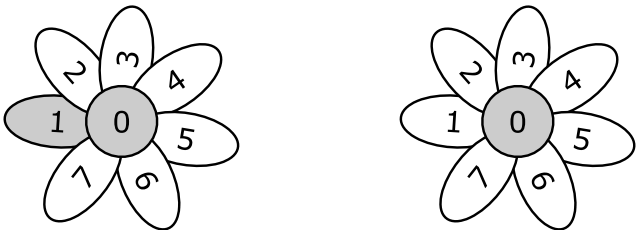

Figure 2: The drawing depicts the concepts $\{0,1\}$ and $\{0\}$ in SF_7. Each concept in SF_n consists of the kernel 0 and at most one of the petals $1,\ldots,n$. The class is named after the sunflower and fulfills the well known definition of "sunflower" in combinatorics, but is otherwise unrelated.

We would like to compare our variant of the disagreement coefficient with Hanneke's definition from [12]. Let $B(h,r)$ denote the closed ball of radius r around h in $\mathcal{C}$ and let $\theta_H(\mathcal{C}|\mathbb{P},h^*)$ denote Hanneke's disagreement coefficient:

$$\begin{aligned} B(h^*,r) &= \{h \in \mathcal{C} | \mathbb{P}(h \neq h^*) \leq r\} \\ \theta_H(\mathcal{C}|\mathbb{P},h^*) &= \sup_{r>0} \frac{\mathbb{P}(\mathrm{DIS}(B(h^*,r)))}{r} \end{aligned}$$

Let $r^* := \sup_{h \in V_{\mathcal{C}}(X',h^*)} \mathbb{P}(h \neq h^*)$. Obviously, the version space is contained in a ball of radius r^*, thus:

$$\theta(\mathcal{C},\mathcal{C}|\mathbb{P},X',h^*) = \frac{\mathbb{P}(\mathrm{DIS}(V_{\mathcal{C}}(X',h^*)))}{r^*} \leq \frac{\mathbb{P}(\mathrm{DIS}(B(h^*,r^*)))}{r^*} \leq \theta_H(\mathcal{C}|\mathbb{P},h^*)$$

Please note that the gap can be arbitrarily large: if $\mathbb{P}$ is the uniform distribution over a finite set X, it is easy to see that $\theta_H(2^X|\mathbb{P},\emptyset) = |X|$, while $\theta(2^X) \leq 2$, as we show in Example 2.

We will now calculate θ for the following class, which will be useful for proving lower bounds in section 4.2:

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

See Figure 2 for a visualization SF_n.

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

Proof. Let $\mathbb{P}$ be uniform on $\{1,\ldots,n\}$, let $X' = h^* = \{0\}$. Then $V := V_{\mathrm{SF}_n}(X',h^*) = \mathrm{SF}_n$ and $\mathrm{DIS}(V) = \{1,\ldots,n\}$ 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 the classical sample size bound of Theorem 2, there exists a sample size $m = \tilde{O}(\text{VCdim}(\mathcal{H})/\varepsilon)$ 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 is summarized in the following

Lemma 2. *There exists a sample size $m = \tilde{O}(\text{VCdim}(\mathcal{H})/\varepsilon)$ 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, $\mathbb{P}((DIS(V_{\mathcal{C}}(X', h^*))) \leq \theta(\mathcal{C}, \mathcal{H}) \cdot \varepsilon$.*

This lemma indicates that one should choose $\mathcal{H}$ so as to minimize $\theta(\mathcal{C}, \mathcal{H}) \cdot \text{VCdim}(\mathcal{H})$. Note that making $\mathcal{H}$ more powerful leads to smaller values of $\theta(\mathcal{C}, \mathcal{H})$ but comes at the price 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 $x \in \text{DIS}(V_{\mathcal{C}}(X', h^*))$. 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.). See also Theorem 4 and its proof.

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 $\text{DIS}(V_{\mathcal{C}}(X', h^*)) \subseteq \{x |\ h_{min}(x) \neq h_{max}(x)\}$. From this and the triangle inequality, it follows that

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

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

Example 2. *Since POWERSET, the class consisting of all subsets of a finite set X, and HALFINTERVALS, the class consisting of sets of the form $(-\infty, a)$ with $a \in \mathbb{R}$, are closed under intersection and union, we obtain $\theta(\text{POWERSET}) \leq 2$ and $\theta(\text{HALFINTERVALS}) \leq 2$.*

Let the class $\mathcal{C}$ consist of both the open and the closed homogeneous half planes and let $\mathcal{H}$ be the class of unions and intersections of two half planes from $\mathcal{C}$. It is easy to see that $\mathcal{H}$ contains hypotheses with plus-sided errors (the smallest pie slice with apex at $\vec{0}$ that includes all positive examples in a sample; see Figure 3) and hypotheses with minus-sided errors (the complement of the smallest pie slice with apex at $\vec{0}$ 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 at the Disagreement Coefficient

In Section 3.1 we investigate the question how small the product $\text{VCdim}(\mathcal{C}) \cdot \theta(\mathcal{C}, \mathcal{H})$ 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

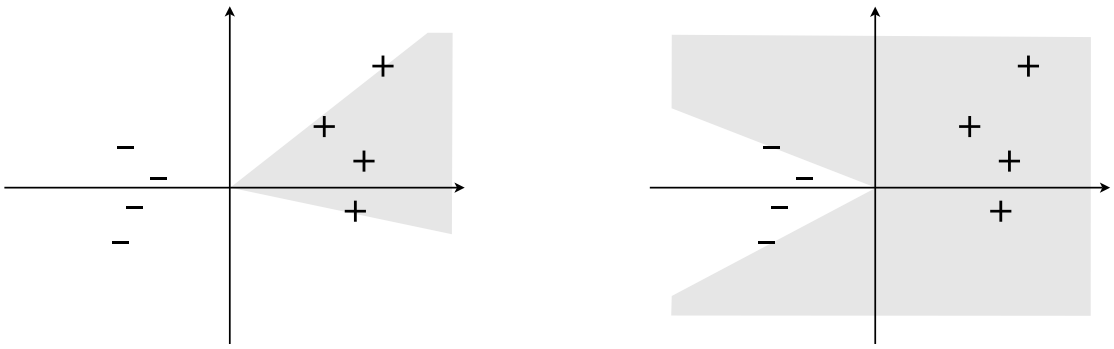

Figure 3: The shaded areas represent hypotheses with plus- and minus-sided errors. Note that the disagreement region, as seen in Figure 1, lies inside of the minus-sided hypothesis and outside of the plus-sided one.

of the largest singleton subclass[3] of $\mathcal{C}$. If $\mathcal{C}$ contains singleton subclasses of arbitrary size, we define $s^+(\mathcal{C})$ as infinite. 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 $\theta(\mathcal{C}, \mathcal{C}^+) \leq 2$ so that* $\mathrm{VCdim}(\mathcal{C}^+) \cdot \theta(\mathcal{C}, \mathcal{C}^+) \leq 2s^+(\mathcal{C})$.

Proof. By construction, $\mathcal{C} \subseteq \mathcal{C}^+$ and $\mathcal{C}^+$ is closed under union. From this it follows that $s^+(\mathcal{C}) \leq \mathrm{VCdim}(\mathcal{C}^+)$. 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 $\mathrm{VCdim}(\mathcal{C}^+) \leq s^+(\mathcal{C})$. 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, $\theta(\mathcal{C}, \mathcal{C}^+) \leq 2$. □

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[4] of $\mathcal{C}$. If $\mathcal{C}$ contains co-singleton subclasses of arbitrary size, we define $s^-(\mathcal{C})$ as infinite. 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 $\theta(\mathcal{C}, \mathcal{C}^-) \leq 2$ so that* $\mathrm{VCdim}(\mathcal{C}^-) \cdot \theta(\mathcal{C}, \mathcal{C}^-) \leq 2s^-(\mathcal{C})$.

We now arrive at the following general bound:

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

[3] A singleton subclass of $\mathcal{C}$ is a set $\mathcal{C}' \subseteq \mathcal{C}$ such that for each $h \in \mathcal{C}'$ there exists an $x \in h$ that is not contained in any other $h' \in \mathcal{C}'$.

[4] A co-singleton subclass of $\mathcal{C}$ is a set $\mathcal{C}' \subseteq \mathcal{C}$ such that for each $h \in \mathcal{C}'$ there exists an $x \notin h$ that is contained in every other $h' \in \mathcal{C}'$.

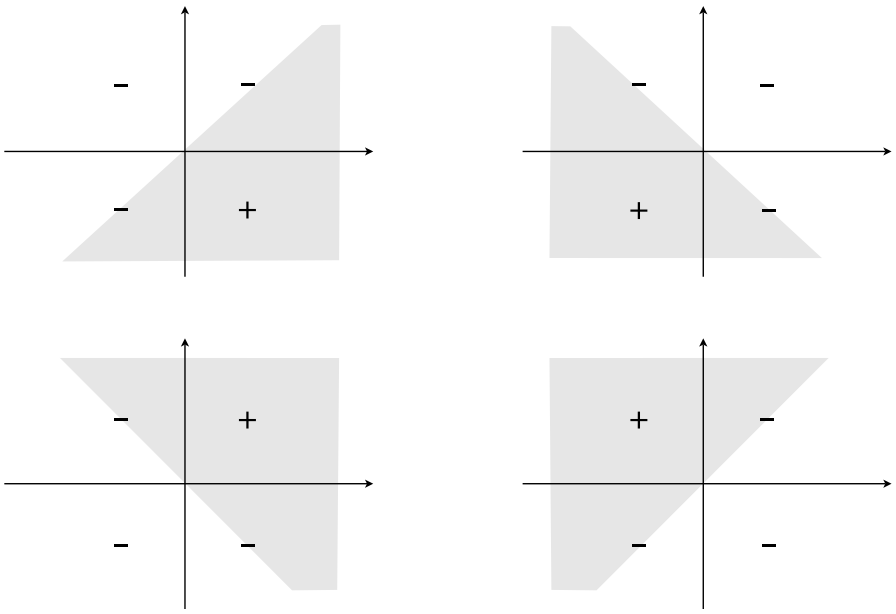

Figure 4: Using four fixed points, one can find the singletons of size four as a subclass of the open, homogeneous half planes. A co-singleton class of size four is induced by the complementary, closed half planes.

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

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

is the upper bound from Sauer's Lemma [14]. 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 $\mathrm{VCdim}(\mathcal{H}) \leq 2\max\{s^+(\mathcal{C}), s^-(\mathcal{C})\}$. Finally note that $\theta(\mathcal{C}, \mathcal{H}) \leq 2$ 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 [10] as the "unique negative dimension" of $\mathcal{C}$. He showed that it characterizes PAC-learnability from positive examples alone.

Example 3. *Let us continue with the classes from Example 2. As noted before, POWERSET and HALFINTERVALS are closed under union and intersection, and we yield $\mathcal{C}^+ = \mathcal{C}^- = \mathcal{C}$ for both classes. Yet they differ strongly in their singleton sizes: the power set over n elements contains both a singleton and a co-singleton class of size n, so we have $s^+(POWERSET) = s^-(POWERSET) = n$, while the class of half intervals only satisfies $s^+(HALFINTERVALS) = s^-(HALFINTERVALS) = 1$. The latter is due to the fact that for any two points on the real line no half interval can assign a negative label to the lower point and a positive label to the higher one simultaneously.*

Now, let $\mathcal{C}$ denote the class of both the open and the closed homogeneous half planes again. We have already seen the classes $\mathcal{C}^-$ and $\mathcal{C}^+$ in Example 2: clearly, $\mathcal{C}^-$ consists of all open and closed pie slices with apex $\vec{0}$ and $\mathcal{C}^+$ consists of the complements of such pie slices (see Figure 3). It is also easy to see that both s^+ and s^- are at least four (see Figure 4). We can see that four is also an upper bound of s^+ (analogous for s^-), because for any choice of five half planes each of which contains at least one of five previously fixed points it holds that one of the points is contained by at least two of the half planes.

Let us conclude with two new examples: the class SF_n *and the class INTERVALS, which consists of the closed and open intervals over* $\mathbb{R}$.

Since SF_n *is closed under intersection, we yield* $\mathcal{C}^-(\mathrm{SF}_n) = \mathrm{SF}_n$. *On the other hand, constructing all possible unions results in the power set over* $\{1,\ldots,n\}$, *thus* $\mathcal{C}^+(\mathrm{SF}_n) = \{\{0\} \cup S | S \subseteq \{1,\ldots,n\}\}$. *This stark difference is also reflected in the singleton and co-singleton sizes: because the elements* $\{1,\ldots,n\}$ *form a singleton class of size* n, *the singleton size* $s^+(\mathrm{SF}_n)$ *is* n, *while the co-singleton size* $s^-(\mathrm{SF}_n)$ *is just one.*

The INTERVALS are even more extreme in this regard. This class is also closed under intersection, thus $\mathcal{C}^- = INTERVALS$, *and the co-singleton size* s^- *is obviously just two. However, the set of all unions is a intricate set of infinite VC-dimension and, because each element in the set* $\mathbb{N}$ *can be covered solitarily by a small interval, the singleton size* s^+ *of the INTERVALS is also infinite.*

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* $k \geq 1$*:* $\theta(\mathcal{C}^{[k]}, \mathcal{H}^{[k]}) = \theta(\mathcal{C}, \mathcal{H})$.

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,

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

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 of seeing a positive example of h^*. Similarly, $p_- = \mathbb{P}(h^* = 0)$ denotes the probability of 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

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

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.

According to the Conditional Independence Assumption, a pair (x_1, x_2) for the learner is generated at random as follows:

1. With probability p_+ commit to a positive example, and with probability $p_- = 1 - p_+$ commit to a negative example of h^*.
2. Conditioned to "+", (x_1, x_2) is chosen at random according to $\mathbb{P}(\cdot|+) \times \mathbb{P}(\cdot|+)$. Conditioned to "−", (x_1, x_2) is chosen at random according to $\mathbb{P}(\cdot|-) \times \mathbb{P}(\cdot|-)$.

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:

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.)

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 behavior 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.

Theorem 5. *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:

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if rule R1 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \max\left\{\frac{1}{p_{min}}, \theta_{max}\right\}}\right) & \text{if rule R2 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1 \theta_2}\right) \text{ resp. } \tilde{O}\left(\sqrt{\frac{\max\{s_1^+ s_2^+, s_1^- s_2^-\}}{\varepsilon}}\right) & \text{if rule R3 is applied} \end{cases} \tag{3}$$

PROOF. By an application of Chernoff-bounds, $\tilde{O}(1)$ 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 to rule R1. Note that the sample size specified in (3) is, by Theorem 2, sufficient to bound (with high probability) the error rate of hypotheses h_1, h_2, respectively, as follows:

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

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. As for rule R2, the proof proceeds analogously. We may assume that (with high probability) the error rate of hypotheses h_1, h_2, respectively, is bounded as follows:

$$\varepsilon_1 = \sqrt{\frac{d_1}{2d_2} \cdot \min\left\{p_{min}, \frac{1}{8\theta_{max}}\right\} \cdot \varepsilon} \text{ and } \varepsilon_2 = \sqrt{\frac{d_2}{2d_1} \cdot \min\left\{p_{min}, \frac{1}{8\theta_{max}}\right\} \cdot \varepsilon}$$

If R2 assigns a wrong label to (x_1, x_2), then $(h_1(x_1) = h_2(x_2) = 0 \wedge h^*(x_1, x_2) = 1) \vee (x_1 \in \mathrm{DIS}_1 \wedge h_2(x_2) = 1 \wedge h^*(x_1, x_2) = 0) \vee (x_2 \in \mathrm{DIS}_2 \wedge h_1(x_1) = 1 \wedge h^*(x_1, x_2) = 0)$. Thus, the error rate induced by R2 is bounded (with high probability) as follows:

$$\begin{aligned} & \mathbb{P}(h_1(x_1) = h_2(x_2) = 0 | +) p_+ + \mathbb{P}(x_1 \in \mathrm{DIS}_1 \wedge h_2(x_2) = 1 | -) p_- \\ & + \mathbb{P}(x_2 \in \mathrm{DIS}_2 \wedge h_1(x_1) = 1 | -) p_- \\ \leq \; & \tfrac{1}{p_+} \cdot \underbrace{\mathbb{P}(h_1(x_1) = 0 | +) p_+}_{\leq \varepsilon_1} \cdot \underbrace{\mathbb{P}(h_2(x_2) = 0 | +) p_+}_{\leq \varepsilon_2} \\ & + \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \Big(\underbrace{\mathbb{P}(x_1 \in \mathrm{DIS}_1 | -) p_-}_{\leq \theta_1 \varepsilon_1} \cdot \underbrace{\mathbb{P}(h_2(x_2) = 1 | -) p_-}_{\leq \varepsilon_2} \\ & + \underbrace{\mathbb{P}(x_2 \in \mathrm{DIS}_2 | -) p_-}_{\leq \theta_2 \varepsilon_2} \cdot \underbrace{\mathbb{P}(h_1(x_1) = 1 | -) p_-}_{\leq \varepsilon_1} \Big) \\ \leq \; & \tfrac{1}{p_{min}} \cdot \varepsilon_1 \varepsilon_2 + 4 \varepsilon_1 \varepsilon_2 \cdot (\theta_1 + \theta_2) \\ \leq \; & \left(\tfrac{1}{p_{min}} + 8\theta_{max}\right) \cdot \varepsilon_1 \varepsilon_2 \\ \leq \; & 2 \cdot \max\left\{\tfrac{1}{p_{min}}, 8\theta_{max}\right\} \cdot \varepsilon_1 \varepsilon_2 \leq \varepsilon \end{aligned}$$

As for rule R3, sample size $\tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1 \theta_2}\right)$ is sufficient to bound (with high probability) the error rate of h_1, h_2, respectively, as follows:

$$\varepsilon_1 = \frac{1}{2} \cdot \sqrt{\frac{d_1}{d_2} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon} \text{ and } \varepsilon_2 = \frac{1}{2} \cdot \sqrt{\frac{d_2}{d_1} \cdot \frac{1}{\theta_1 \theta_2} \cdot \varepsilon}$$

If R3 assigns a wrong label to (x_1, x_2), then $x_1 \in \mathrm{DIS}_1$, $x_2 \in \mathrm{DIS}_2$, and the true label is "$-$". Thus the error rate induced by R3 is bounded (with high probability) as follows:

$$\begin{aligned} \mathbb{P}(x_1 \in \mathrm{DIS}_1 \wedge x_2 \in \mathrm{DIS}_2 | -) p_- &= \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{\mathbb{P}(x_1 \in \mathrm{DIS}_1 | -) p_-}_{\leq \theta_1 \varepsilon_1} \cdot \underbrace{\mathbb{P}(x_2 \in \mathrm{DIS}_2 | -) p_-}_{\leq \theta_2 \varepsilon_2} \\ &\leq 4 \theta_1 \theta_2 \varepsilon_1 \varepsilon_2 \leq \varepsilon \end{aligned}$$

There is an alternative analysis for rule R3 which proceeds as follows. Since we have assumed that $\hat{p}_- \geq 1/2$, R3 assigns label "$+$" to every instance $x_1 \in \mathrm{DIS}_1$ (resp. to every instance $x_2 \in \mathrm{DIS}_2$). We can also achieve this behavior by choosing h_1 (resp. h_2) as the union of all hypotheses from the version space and assigning the label "$+$" to (x_1, x_2) exactly if both h_1 and h_2 agree on a positive label. Recall that, according to our definition of $\mathcal{C}_1^+, \mathcal{C}_2^+$, $h_1 \in \mathcal{C}_1^+$ and $h_2 \in \mathcal{C}_2^+$. Recall furthermore that $\mathrm{VCdim}(\mathcal{C}_1^+) = s_1^+$ and $\mathrm{VCdim}(\mathcal{C}_2^+) = s_2^+$. Thus, sample size $\tilde{O}\left(\sqrt{\frac{s_1^+ s_2^+}{\varepsilon}}\right)$ is sufficient to bound (with high probability) the error rate of h_1, h_2, respectively, as follows:

$$\varepsilon_1 = \frac{1}{2} \cdot \sqrt{\frac{s_1^+}{s_2^+} \cdot \varepsilon} \text{ and } \varepsilon_2 = \frac{1}{2} \cdot \sqrt{\frac{s_2^+}{s_1^+} \cdot \varepsilon}$$

An error of rule R3 can occur only when h_1 errs on x_1 and h_2 errs on x_2, which implies that the true label of (x_1, x_2) is "$-$". According to conditional independence, the probability for this to happen is bounded as follows:

$$\begin{aligned} \mathbb{P}(h_1 \text{ errs on } x_1 \text{ and } h_2 \text{ errs on } x_2) &= \mathbb{P}(h_1 \text{ errs on } x_1 \text{ and } h_2 \text{ errs on } x_2 | -) p_- \\ &= \frac{1}{p_-} \cdot \mathbb{P}(h_1 \text{ errs on } x_1 | p_-) p_- \cdot \mathbb{P}(h_2 \text{ errs on } x_2 | p_-) p_- \\ &= \frac{1}{p_-} \cdot \mathbb{P}(h_1 \text{ errs on } x_1) \cdot \mathbb{P}(h_2 \text{ errs on } x_2) \\ &\leq 4 \cdot \varepsilon_1 \cdot \varepsilon_2 = \varepsilon \end{aligned}$$

By symmetry, assuming that $\hat{p}_- < 1/2$, the upper bound $\tilde{O}\big(\sqrt{s_1^- s_2^- / \varepsilon}\big)$ on the sample size holds. Thus the bound $\tilde{O}\left(\sqrt{\frac{\max\{s_1^+ s_2^+, s_1^- s_2^-\}}{\varepsilon}}\right)$ takes care of all values for $\hat{p}_-$. This concludes the proof of the theorem. □

Please observe that the second upper bound for rule R3 is given completely in combinatorial parameters. Also note that, in the case $\hat{p}_- < 1/2$, finding $h_1 \in \mathcal{C}_1^-$ (resp. $h_2 \in \mathcal{C}_2^-$), which is the smallest consistent hypothesis in $\mathcal{C}_1$, is possible using negative examples only. This shows a strong connection to the results by Geréb-Graus in [10], 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.

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 $\varepsilon > 2/(\theta_1 \theta_2)$ and $\hat{p}_+ < \varepsilon/2$ (or $\hat{p}_- < \varepsilon/2$, resp.), it votes for label "$-$" (or for label "$+$", resp.). If $\varepsilon \leq 2/(\theta_1 \theta_2)$ or $\hat{p}_{min} := \min\{\hat{p}_+, \hat{p}_-\} \geq \varepsilon/2$, then it applies the rule

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

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

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if } \frac{\theta_{min}}{\theta_{max}} \leq p_{min} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right) & \text{if } \frac{1}{\theta_{max}} \leq p_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{1}{p_{min}}}\right) & \text{if } \frac{1}{\theta_1\theta_2} \leq p_{min} < \frac{1}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1\theta_2}\right) \text{ resp. } \tilde{O}\left(\sqrt{\frac{\max\{s_1^+ s_2^+, s_1^- s_2^-\}}{\varepsilon}}\right) & \text{if } p_{min} < \frac{1}{\theta_1\theta_2} \end{cases} \tag{5}$$

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

Proof. We first would like to note that, according to (5), we have at least

$$\tilde{O}\left(\sqrt{\frac{\min\{\theta_1\theta_2, 1/p_{min}\}}{\varepsilon}}\right) \tag{6}$$

labeled examples at our disposal. Furthermore, (5) is a continuous function in p_{min} (even for $p_{min} = \theta_{min}/\theta_{max}, 1/\theta_{max}, 1/(\theta_1\theta_2)$). We proceed by case analysis:

Case 1: $4/\varepsilon < \min\{2\theta_1\theta_2, 1/p_{min}\}$ so that $p_{min} < \varepsilon/4$ and $\varepsilon > 2/(\theta_1\theta_2)$.
Then (6) is at least $\tilde{O}(1/\varepsilon)$ in order of magnitude. We can apply Chernoff-bounds and conclude that, with high probability, $\hat{p}_{min} < \varepsilon/2$. But then the Combined Rule outputs the empirically more likely label, which leads to error rate $p_{min} \leq \varepsilon/4$.

Case 2: $2\theta_1\theta_2 \leq \min\{4/\varepsilon, 1/p_{min}\}$ so that $p_{min} \leq 1/(2\theta_1\theta_2)$ and $\varepsilon \leq 2/(\theta_1\theta_2)$.
Then, (6) is at least $\tilde{O}(\theta_1\theta_2)$. We can apply Chernoff-bounds and conclude that, with high probability, $\hat{p}_{min} \leq \frac{1}{\theta_1\theta_2}$. But then rule R3 is applied which, according to Theorem 5, leads to the desired upper bound on the sample size.

Case 3: $1/p_{min} \leq \min\{4/\varepsilon, 2\theta_1\theta_2\}$.
We can apply Chernoff-bounds and conclude that, with high probability, $\hat{p}_{min}$ and p_{min} differ by factor 2 only. If the Combined Rule outputs the empirically more likely label, then $\hat{p}_{min} < \varepsilon/2$ and, therefore, the resulting error rate p_{min} is bounded by ε. Let us now assume that $\hat{p}_{min} \geq \varepsilon/2$ so that the Combined Rule proceeds according to (4). If the learner could substitute the (unknown) p_{min} for $\hat{p}_{min}$ within (4), we could apply Theorem 5 and would be done. But since, as mentioned above, (5) is a *continuous* function in p_{min}, even the knowledge of $\hat{p}_{min}$ is sufficient. □

4.2. Lower Bounds on the Sample Size

4.2.1. A lower bound archetype

In this section, we prove a lower bound on the sample complexity for the class SF_n from Lemma 1. Note that all lower bounds obtained for SF_n immediately generalize to concept classes containing SF_n as subclass.

All other lower bounds in this paper apply Lemma 6 directly or use the same proof technique.

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 "+" be the (a-priori) less likely label, i.e., $p_+ = p_{min}$, let $\mathcal{C}_1$ be a concept class over domain $X_1 = \{a_0, a_1, \ldots, a_{n_1+2}\}$ (with a_0 in the role of the center-point belonging to every concept from $\mathcal{C}_1$), and let $\mathcal{C}_2$ be a concept class over domain $X_2 = \{b_0, b_1, \ldots, b_{n_2+2}\}$, respectively. Let $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Obviously,

$$p_+ = p_{min} \leq \frac{1}{(n_1+2)(n_2+2)} = \frac{1}{\theta_1\theta_2} \ . \tag{7}$$

Consider the following malign scenario:

- Index s is uniformly chosen at random from $X_1 \setminus \{0,1\}$. It represents a randomly chosen target concept $h_1^* = \{a_0, a_s\} \in \mathcal{C}_1$. Similarly, index t is uniformly chosen at random from $X_2 \setminus \{0,1\}$ and represents a randomly chosen target concept $h_2^* = \{b_0, b_t\} \in \mathcal{C}_2$. We define $\mathbb{P}(a_s|+) = \sqrt{\varepsilon_1/p_+}$ and $\mathbb{P}(b_t|+) = \sqrt{\varepsilon_2/p_+}$. In the sequel, points a_s and b_t are called "secret": if one of them occurs in the sample, the learner will have enough knowledge to be error-free on test-points. Note that the secret points have a high chance to remain hidden from the learner (i.e., to not occur in the sample) but still have too much probability mass for being neglected.

- $\mathbb{P}(a_0|+) = 1 - \mathbb{P}(a_s|+)$ and $\mathbb{P}(b_0|+) = 1 - \mathbb{P}(b_t|+)$. Note that a_0, a positive example for every concept in $\mathcal{C}_1$, is a redundant instance that absorbs a high fraction of the total probability mass. The analogous remark applies to b_0.

- $\mathbb{P}(a_1|-) = 1 - 4 \cdot \sqrt{n_1\varepsilon_1/n_2}$ and $\mathbb{P}(b_1|-) = 1 - 4 \cdot \sqrt{n_2\varepsilon_2/n_1}$. The effect of assigning much probability mass to a_1 and b_1 is that many negative examples different from a_1 or b_1 will not find their way into the sample.

- The instances from $X_1 \setminus \{a_0, a_1, a_s\}$ evenly share a minus-conditional probability mass of $1 - \mathbb{P}(a_1|-)$. The instances from $X_2 \setminus \{b_0, b_1, b_t\}$ evenly share a minus-conditional probability mass of $1 - \mathbb{P}(b_1|-)$.

Let us assume that the sample size satisfies $m \leq m_0$ for

$$m_0 = \frac{\sqrt{n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon} \ .$$

We will show that, with a probability of at least $1/2$, an error rate of at least ε is unavoidable. To this end, we proceed as follows: Let Z_1 count the number of sample points that hit $X_1 \setminus \{a_0, a_1, a_s\}$ (the "interesting points" in X_1, besides a_s) and let Z_2 count the number of sample points that hit $X_2 \setminus \{b_0, b_1, b_t\}$ (the "interesting points" in X_2, besides b_t). Then the following holds:

- We can bound the expectation of Z_b for $b = 1, 2$ (recall that $\varepsilon_1 = \varepsilon_2 = \varepsilon$):

$$\mathbb{E}[Z_b] \leq (1 - p_+) \cdot 4 \cdot \sqrt{n_b\varepsilon_b/n_{3-b}} \cdot \frac{\sqrt{n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon} \leq \frac{n_b}{10}$$

  By the Markov-inequality it follows that the probability that $Z_b > n_1/2$ is at most $p_b = 1/5$.

- Using Equation (7) the number of occurrences of a_s (or b_t, resp.) in the sample can be upper bounded by

$$p_+ \cdot \sqrt{\varepsilon_1/p_+} \cdot \frac{\sqrt{n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon} = \sqrt{n_1 n_2 p_+}/40 \leq 1/40 \ .$$

  Consequently with probability $p_3 = 1/40$ a_s does not occur in the sample. The same holds for b_t with probability $p_4 = 1/40$.

Denote by H_1 (resp. H_2) the subset of $X_1 \setminus \{a_0, a_1, a_s\}$ (resp. $X_2 \setminus \{b_0, b_1, b_t\}$) consisting of all those points that do not occur in the sample. According to the above calculations, the probability that H_1 and H_2 contain at least half of the possible points, and neither a_s nor b_t occurs in the sample, is at least $1 - p_1 - p_2 - p_3 - p_4 > 1/2$. In the rest of the proof we assume that this holds.

Before we analyze the Bayes-error (smallest possible error-rate), we assume, to the advantage of the learner, that not only the labeled sample is revealed but also the probabilities p_+, $\mathbb{P}(a_0|+)$, $\mathbb{P}(b_0|+)$, $\mathbb{P}(H_1|-)$, $\mathbb{P}(H_2|-)$, and the fact that s (or t, resp.) was chosen uniformly at random from $\{2, \ldots, n_1 + 2\}$ (or from $\{2, \ldots, n_2 + 2\}$, resp.). 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.

How is the situation for test-instances from U? The crucial observation is that the plus- and the minus-conditional a-posteriori probabilities assign precisely the same value to each pair in U. This might look

wrong at first glance because, for example, $U_1 = H_1 \cup \{a_s\}$ and a_s is the only positive example in U_1. Thus, $\mathbb{P}(a_s|+) = 1 - \mathbb{P}(a_0|+) > 0$ whereas $\mathbb{P}(x_1|+) = 0$ for every $x_1 \in H_1$. But recall that s had been chosen uniformly at random from $\{2, \ldots, n_1 + 2\}$. Thus, the a-posteriori distribution of the random variable a_s (reflecting the perspective of the learner after its evaluation of the labeled random sample) is uniform on U_1. The arguments for X_2 and for negative examples are similar. Thus, by symmetry, the plus- and minus-conditional a-posteriori probabilities assign the same value to every point in U so that the Bayes-decision on instances $(x_1, x_2) \in U$ takes the following mutually equivalent forms:

- Vote for the label with the higher a-posteriori probability given (x_1, x_2).
- Always vote for the label with the higher a-posteriori probability given U (the information that the test-instance belongs to U).
- If $\mathbb{P}(E_+ \cap U) \geq \mathbb{P}(E_- \cap U)$ vote always for "+", otherwise vote always 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

$$\mathbb{P}(U \cap E_-) \geq \underbrace{(1 - p_+)}_{\geq 1/2} \cdot (2\sqrt{n_1\varepsilon_1/n_2}) \cdot (2\sqrt{n_2\varepsilon_2/n_1}) \geq 2\varepsilon \ .$$

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

The following corollary demonstrates the use of the padding argument to boost lower bounds of the type used in Lemma 6 to classes with arbitrary VC-dimensions:

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. $\theta(\mathcal{C}_b) = \theta(\mathcal{C}_b^{[d_b]})$ follows from Lemma 5.

A malicious scenario for the classes $\mathcal{C}_b^{[d_b]}$ is obtained by installing the malicious scenario from the proof of Lemma 6 (with minor modifications that will be explained below) for each of the d_1 many copies of $\mathcal{C}_1, X_1$ and for each of the d_2 many copies of $\mathcal{C}_2, X_2$:

- The role of the secret point a_s is now played by the secret points $(a_{s(i)}, i)$ for $i = 1, \ldots, d_1$. Here, $s(i)$ is uniformly chosen at random from $\{2, \ldots, n_1 + 2\}$. The analogous remark applies to b_t.
- Instead of setting $\varepsilon_1 = \varepsilon_2 = \varepsilon$, we set

$$\varepsilon_1 = \sqrt{\frac{d_1}{d_2}} \cdot \varepsilon \text{ and } \varepsilon_2 = \sqrt{\frac{d_2}{d_1}} \cdot \varepsilon \ .$$

- The plus- and minus-conditional probabilities for elements of $X_1^{(i)}$, $i = 1, \ldots, d_1$, are given by the same formulas as the probabilities for the corresponding elements of X_1 except for a scaling factor of $1/d_1$ (for reasons of normalization). But note that these formulas are given in terms of (the redefined) ε_1. The analogous remark applies to the plus- and minus-conditional probabilities for elements of $X_2^{(j)}$, $j = 1, \ldots, d_2$.

Let us assume that the sample size satisfies

$$m \leq \frac{\sqrt{d_1 d_2 n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon}$$

In comparison to Lemma 6, the sample size is scaled-up by factor $\sqrt{d_1 d_2}$. We will show that, with a probability of at least 1/2, an error rate of at least $\varepsilon/4$ is unavoidable (which, after consistently replacing ε by 4ε, would settle the proof for the theorem). To this end, we proceed as follows:

- $Z_1^{(i)}$ counts the number of sample points that hit $X_1^{(i)} \setminus \{(a_0, i), (a_1, i), (a_{s(i)}, i)\}$, and $Z_1^{[d_1]} = \sum_{i=1}^{d_1} Z_1^{(i)}$. The hitting probability (except for scaling-factor $1/d_1$ and the redefinition of ε_1 the same as in the proof of Lemma 6) is

$$\frac{1}{d_1}(1-p_+) \cdot 4 \cdot \sqrt{\frac{n_1 \varepsilon_1}{n_2}} = \frac{1}{\sqrt{d_1 d_2}}(1-p_+) \cdot 4 \cdot \sqrt{\frac{n_1 \varepsilon}{n_2}} \ .$$

  The number of trials is m. Thus,

$$\mathbb{E}[Z_1^{(i)}] = \frac{1}{\sqrt{d_1 d_2}}(1-p_+) \cdot 4 \cdot \sqrt{\frac{n_1 \varepsilon}{n_2}} \cdot m \leq \frac{n_1}{10} \ .$$

  In other words, $\mathbb{E}[Z_1^{(i)}]$ has the same upper bound as $\mathbb{E}[Z_1]$. We may now conclude that $\mathbb{E}[Z_1^{[d_1]}] \leq d_1 n_1/10$. Thus, with a probability of at least $1 - 1/5$, $Z_1^{[d_1]} \leq d_1 n_1/2$, which is assumed in the sequel. Note that $d_1 n_1$ is the number of interesting negative examples in $X_1^{[d_1]}$. Thus at least half of them remain hidden from the learner. Their total minus-conditional probability mass is therefore at least $2 \cdot \sqrt{n_1 \varepsilon_1 / n_2}$ (the same as in the proof of Lemma 6).

- The fact that $\mathbb{E}[Z_1^{(i)}]$ has the same upper bound as $\mathbb{E}[Z_1]$ is no accident: compared to Lemma 6, the upper bound on the sample size scales-up by factor $\sqrt{d_1 d_2}$ and the hitting probabilities for events in $X^{(i)}$ scale-down by the same factor. For this reason, we obtain similar considerations for random variable Z_2 and may conclude that, with a probability of at least $1-1/5$, half of the interesting negative examples in $X_2^{[d_2]}$ remain hidden from the learner. Their total negative-conditional probability mass is therefore at least $2 \cdot \sqrt{n_2 \varepsilon_2 / n_1}$ (the same as in the proof of Lemma 6).

- Similarly, we get that the expected number of sample points that hit $\{a_{s(i)}^{(i)} |\ i = 1, \ldots, d_1\}$ is bounded by $d_1/40$. Thus, with a probability of at least $1-1/20$, at least half of the secret points in $X_1^{[d_1]}$ remain hidden from the learner. Their total plus-conditional probability mass is at least $1/2 \cdot \sqrt{\varepsilon_1 / p_+}$ (half of the probability mass of the secret point a_s in the proof of Lemma 6). Analogously with probability at least $1 - 1/20$, at least half of the secret points in $X_2^{[d_2]}$ remain hidden from the learner. Their total plus-conditional probability mass is therefore at least $1/2 \cdot \sqrt{\varepsilon_2 / p_+}$.

Let U denote the set of test-instances from $X_1^{[d_1]} \times X_2^{[d_2]}$ such that none of its two components occurred in the sample. By symmetry (as in the proof of Lemma 6), the plus- and minus-conditional probabilities assign the same value to any point in U, respectively. Let E_+ (or E_-, resp.) denote the set of test-instances which are labeled "+" (or labeled "−", resp.). As in in the proof of Lemma 6, the Bayes-error is given by $\min\{\mathbb{P}(E_+ \cap U), \mathbb{P}(E_- \cap U)\}$, and an easy calculation (similar to the calculation in in the proof of Lemma 6) shows that it is bounded from below by $\varepsilon/4$.[5] □

[5] Factor 4 is lost in comparison to Lemma 6 because now half of the secret points in $X_1^{[d_1]}$ and $X_2^{[d_2]}$, respectively, might perhaps occur in the sample whereas, in the proof of Lemma 6, we could assume that a_s and b_t remained hidden from the learner.

4.2.2. A purely combinatorial lower bound

Here comes the lower bound that is tight from the perspective of a worst case analysis, which is making use of rather small values of p_{min}. It matches nicely with the combinatorial upper bound from Theorem 5:

Theorem 6. *Assume that $3 \leq s_b^+, s_b^- < \infty$ 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 $\Omega\big(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}\big)$.*

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 $n_b + 2$. □

One can drop the restriction "$3 \leq s_b^+, s_b^-$" and still prove tight bounds. The corresponding results, which need a tedious case distinction, are to be published in a follow-up paper.

Note that Theorem 6 implies a weak converse of Theorem 4 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 6 would exceed the first upper bound for rule R3 in Theorem 5 with $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$.

4.2.3. Matching lower bounds for Corollary 2

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 worst case. Actually, Corollary 3 is a first step in this direction because the lower bound in this result matches with the upper bound in Corollary 2 when $p_{min} \leq 1/(\theta_1\theta_2)$. We list here some more results of this kind which together witness that all upper bounds mentioned in Corollary 2 are fairly tight. The proof technique is the same as for the "archetypical" lower bounds from section 4.2.1.

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 $\theta(\mathcal{C}) = \theta(\mathrm{co}(\mathcal{C}))$.

Theorem 7. *For sufficiently small $\varepsilon > 0$ at least*

$$\begin{cases} \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{for } \mathcal{C}_1 = \mathrm{SF}_{\theta_1},\ \mathcal{C}_2 = \mathrm{co}(\mathrm{SF}_{\theta_2}) \quad \text{and } \frac{\theta_{min}}{\theta_{max}} \leq p_{min} \\ \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right) & \text{for } \mathcal{C}_1 = \mathrm{SF}_{\theta_1},\ \mathcal{C}_2 = \mathrm{co}(\mathrm{SF}_{\theta_2}) \quad \text{and } \frac{1}{\theta_{max}} \leq p_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{1}{p_{min}}}\right) & \text{for } \mathcal{C}_1 = \mathcal{C}_2 = \mathrm{SF}_{\theta_1} \quad \text{and } \frac{1}{\theta_1\theta_2} \leq p_{min} < \frac{1}{\theta_{max}} \\ \Omega\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1\theta_2}\right) & \text{for } \mathcal{C}_1 = \mathrm{SF}_{\theta_1},\ \mathcal{C}_2 = \mathrm{SF}_{\theta_2} \quad \text{and } p_{min} < \frac{1}{\theta_1\theta_2} \end{cases}$$

many examples are needed to learn $\mathcal{C}_1^{[d_1]}, \mathcal{C}_2^{[d_2]}$ in the PAC Co-training Model under the Conditional Independence Assumption.

PROOF. Note that the lower bound $\Omega\left(d_1 d_2 \theta_1 \theta_2/\epsilon\right)$ is already proved in Corollary 3. The table in Appendix A contains the necessary modifications to the proof of Lemma 6 to obtain the rest of the desired results in case $d_1 = d_2 = 1$. The general versions then can be proved from these along the line of Corollary 3. □

Notice that the lower bounds in Theorem 6 and 7 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 $\mathcal{H}_b = \mathcal{C}_b^+ \cup \mathcal{C}_b^-$.

Theorem 8. *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.

PROOF. If $\varepsilon > 4/(\theta_1\theta_2)$ and $\hat{p}_{min} \leq \varepsilon/2$, the learner will output the less likely label (and this is analyzed as in the proof of Corollary 2). Assume now that $\varepsilon \leq 4/(\theta_1\theta_2)$ or $\hat{p}_{min} > \varepsilon/2$. For reasons of symmetry, we may assume that, for $b = 1, 2$, $\mathcal{H}_b$ contains hypotheses with plus-sided errors w.r.t. $\mathcal{C}_b$. Let h_1, h_2 be two hypotheses that err on positive examples only. Thus, if $h(x_1) = 1$ or $h(x_2) = 1$, the learner may safely vote for label "+". Assume that $h(x_1) = h(x_2) = 0$. If $\hat{p}_{min} \geq 1/(\theta_1\theta_2)$ (Case 1), the learner votes for label "−". Otherwise (Case 2), the learner applies rule R3. According to Theorem 5. R3 leads to the sample size bound $\tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1\theta_2}\right)$.[6] So we may focus on Case 1. The sample size $\tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{1}{p_{min}}}\right)$ is sufficient to bound (with high probability) the error rate of h_1, h_2, respectively, as follows:

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

Thus, the error rate for guessing label "−" in Case 1 is bounded as follows:

$$\begin{aligned} \mathbb{P}(h_1(x_1) = h_(x_2) = 0|+)p_+ &\leq \frac{1}{p_+} \cdot \underbrace{\mathbb{P}(h_1(x_1) = 0|+)p_+}_{\leq \varepsilon_1} \cdot \underbrace{\mathbb{P}(h(x_2) = 0|+)p_+}_{\leq \varepsilon_2} \\ &\leq \frac{1}{p_{min}} \cdot \varepsilon_1\varepsilon_2 = \varepsilon \end{aligned}$$

□

Note that the upper bound from Theorem 8 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 third and fourth lower bound from Theorem 7.

Theorem 9. *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.

[6] Recall from the proof of Corollary 2 that knowing the empirical estimate $\hat{p}_{min}$ instead of the true value p_{min} does not cause much trouble.

PROOF. As in the proof of Theorem 8, we may focus on the case where $\varepsilon \leq 4/(\theta_1\theta_2)$ or $\hat{p}_{min} > \varepsilon/2$. Let h_1, h_2 be two hypotheses such that h_1 errs on positive examples only, and h_2 errs on negative examples only. The learner has an error-free decision unless $x_1 \in \mathrm{DIS}_1$ and $x_2 \in \mathrm{DIS}_2$. Note that $x_1 \in \mathrm{DIS}_1$ implies that $h_1(x_1) = 0$, and $x_2 \in \mathrm{DIS}_2$ implies that $h_2(x_2) = 1$. In case of conflict, the learner applies rule R1 if $\hat{p}_{min} \geq \theta_{min}/\theta_{max}$ (Case 1), and rule R2 (voting for the empirically less likely label) otherwise. According to Theorem 5, R1 leads to the sample size bound $\tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right)$. So we may focus on Case 2. As in the proof of Theorem 5, we may assume that, if $p_{min} < 1/4$, then $\hat{p}_{min} < 1/2$. For reasons of symmetry, we may assume furthermore that $\hat{p}_- \geq 1/2$ (so that R2, in case of conflict, makes the learner vote for label "+"). Our assumptions imply that $p_- \geq 1/4$. Note that the sample size $\tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_{max}}\right)$ is sufficient to bound (with high probability) the error rates of h_1, h_2, respectively, as follows:

$$\varepsilon_1 = \sqrt{\frac{d_1}{d_2} \cdot \frac{1}{\theta_{max}} \cdot \varepsilon} \text{ and } \varepsilon_2 = \sqrt{\frac{d_2}{d_1} \cdot \frac{1}{\theta_{max}} \cdot \varepsilon}$$

Thus, the error rate induced by R2 in Case 2 is bounded as follows:

$$\begin{aligned} \mathbb{P}(x_1 \in \mathrm{DIS}_1 \wedge h_2(x_2) = 1|-)p_- &\leq \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{\mathbb{P}(h_1(x_1) = 0|-)p_-}_{\leq \theta_1\varepsilon_1} \cdot \underbrace{\mathbb{P}(h_2(x_2) = 1|-)p_-}_{\leq \varepsilon_2} \\ &\leq 4 \cdot \theta_{max}\varepsilon_1\varepsilon_2 \\ &= \varepsilon \end{aligned}$$

□

Note that the upper bound from Theorem 9 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 first and second lower bound from Theorem 7.

4.4. Co-training with α-expansion

Because the Conditional Independence assumption is very strict and rarely fulfilled in practical problems, there is some effort to replace Conditional Independence with relaxed assumptions that still allow semi-supervised Co-training to save a significant amount of labels.

To this end, a model called "Co-training with α-expansion" was introduced by Balcan, Blum and Yang in [2]. The learner in this framework learns from positive examples alone, and is required to use a hypothesis with plus-sided error (i.e., one that always returns label "−" if there is any doubt about the true label). One motivation from [2] for this restriction stems from the fact that the assumption we have to make on $\mathbb{P}$ is now only necessary to be satisfied on the positively labeled part of the domain, which is said to be a more realistic assumption in practice.

In this model, just as before, two concept classes $\mathcal{C}_1$ and $\mathcal{C}_2$ over domains X_1 resp. X_2 are given, and the domain distribution $\mathbb{P}$, from which the random examples $(x_1, x_2) \in X_1 \times X_2$ are drawn, is perfectly compatible with the target concepts $h_1^* \in \mathcal{C}_1$ and $h_2^* \in \mathcal{C}_2$, meaning that $h_1^*(x_1) = h_2^*(x_2)$ with probability 1. Let D^+ denote the distribution conditioned on positive labels and let X_1^+ and X_2^+ denote the support of the marginal distributions of D^+ over X_1 resp. X_2. We say that D^+ is *α-expanding with respect to hypothesis classes* H_1, H_2 if for all $h_1 \in H_1$ and $h_2 \in H_2$:

$$\mathbb{P}\big(h_1(x_1) \neq h_2(x_2)\big| +\big) \geq \alpha \cdot \min\left\{\mathbb{P}\big(h_1(x_1) = h_2(x_2) = 1\big| +\big), \mathbb{P}\big(h_1(x_1) = h_2(x_2) = 0\big| +\big)\right\}$$

where condition "+" refers to the event $h^*(x_1) = h^*(x_2) = 1$.

The authors of [2] show that the α-expansion assumption is weaker than Conditional Independence and in [1] it is mentioned, that, if both D^+ and D^- are α-expanding (with $\alpha > 0$), a semi-supervised learner

only needs to see just one (positively or negatively) labeled example to learn successfully, just like under the Conditional Independence Assumption.

One can ask whether this weaker assumption is also always helpful for a fully supervised learner. Below we answer this in the negative by showing an example where, in comparison with general PAC-learning, fully supervised Co-training requires significantly less examples under the Conditional Independence Assumption, but not under the α-expansion assumption.

We use the setting of Example 1 from [2]. The two concept classes, which are also the hypothesis classes, are axis-aligned rectangles in $\mathbb{R}^d$, the two target concepts are chosen to be always identical $h_1^* = h_2^* = h^*$ and non-empty. Note that this also implies $X_1^+ = X_2^+$. Let $p_+ = p_- = 1/2$. Distribution D^+ over $X_1^+ \times X_2^+$ is defined as follows.[7] First x_1 is drawn uniformly at random from X_1^+, i is drawn uniformly at random from $\{1, \ldots, d\}$ and then x_2 is identified with x_1 except for its i-th component which is chosen uniformly at random so that x_2 still lies in X_2^+. This distribution is $\Omega(1/d)$-expanding [2].

Theorem 10. *For $d \geq 2$ and $\varepsilon \leq 1/2$ the following holds in the $1/d$-expanding rectangle learning setting described above: any fully supervised PAC-learner needs at least*

$$m = \Omega\left(\frac{d}{\varepsilon}\right)$$

many labeled examples to learn successfully.

Proof. We will reduce learning the pair (h_1^*, h_2^*) to learning the rectangle h^* in the normal, non co-training setting.

To guarantee that the combined hypothesis has a plus-sided error, the learner has to choose h_1 and h_2 as the smallest rectangles consistent with the sample so that it outputs "$-$" whenever both of x_1 and x_2 lie in the disagreement region. However, from the symmetry $h_1^* = h_2^*$ follows that all information in the sample about h_1^* can also be applied to learn h_2^* and vice versa. Thus the learner can always output a hypothesis of the form (h, h) without risking a higher error rate.

Because $p_+ = 1/2$ and the learner is required to use hypothesis with plus-sided error, it suffices to bound the error under distribution D^+. Let h be any rectangle contained in h^*. We denote the error of (h, h) with respect to the distribution D^+ by $\mathrm{err}_{D^+}(h)$. To model the standard learning setting, let U be the uniform distribution over h^* and denote the error of h with respect to U by $\mathrm{err}_U(h)$. We claim that

$$\mathrm{err}_{D^+}(h) \geq \frac{d-1}{d} \cdot \mathrm{err}_U(h) \ . \tag{8}$$

The reason is as follows. Let x_1 be a misclassified point in the normal PAC-learning setting. Then there can be at most one axis along which x_1 can be moved inside the rectangle h. But then (x_1, x_2) can only be classified correctly, if x_1 and x_2 differ exactly on the corresponding component. The probability of the latter event is $1/d$ so that the error probability $\mathrm{err}_U(h)$ is reduced by factor $(d-1)/d$ only.

Observe that one positive training example (x_1, x_2) in the co-training model does not contain more information about the target rectangle than two examples in standard learning where *all* components of $x_2 \in X_2^+$ (not just the i-th component for a randomly chosen i) are chosen uniformly at random.

Because of (8), it suffices to show that the best learner in the standard setting needs $\Omega(d/\varepsilon)$ many examples to reach an error smaller than ε with high probability.

Let us choose a target rectangle h^* with sides of length 1 (i.e., a cube of volume 1). Assume furthermore that $m \leq d/(8\varepsilon)$ and let $x_1, ..., x_m$ be a sample drawn independently from U.

According to the union bound, the probability that the i-th component of at least one of the m examples lies outside of an interval of length $1 - 4\varepsilon/d$ is at most

$$\frac{4\varepsilon}{d} \cdot m \leq \frac{1}{2} \ .$$

[7] The distribution on the negative points is not important.

Thus the probability of the complementary event, that the i-th component of all sample points lies on said interval, is at least $1/2$.

Additionally, the probability that this holds for at least half of the d components is at least $1/2$. But then, with probability at least $1/2$, the error of the smallest consistent rectangle h is at least

$$1 - (1 - 4\varepsilon/d)^{d/2} \geq 1 - e^{-2\varepsilon} \geq \epsilon \ ,$$

where the last inequality holds thanks to $\varepsilon \leq 1/2$. This completes the proof. □

Note that the authors of [3] show a lower bound of $\Omega(1/\varepsilon)$ on the sample size of any learner (even semi-supervised learners) which learns thresholds, and thereby also intervals, under any continuous distribution. Although it looks like one could apply their result in the last part of the proof of Theorem 10, this is not possible since we fixed p_+.

Also observe that the bound from Theorem 10 is of the same order of magnitude as in the case of standard PAC-learning. So the learner doesn't gain any advantage from Co-training with α-expansion in this example. In contrast, we may conclude from Theorem 8 that $\tilde{O}\big(d/\sqrt{\varepsilon}\big)$ labeled examples are sufficient for learning d-dimensional axis-aligned rectangles in the PAC Co-training Model under the Conditional Independence Assumption.

5. Final Remarks and Open Questions

We close this paper with some final remarks and open questions.

The framework with k views. We looked into a generalized framework where the learner is provided with sample points consisting of k-tuples $(x_1, \ldots, x_k)$ instead of pairs and deals with k concept classes $\mathcal{C}_1, \ldots \mathcal{C}_k$ and the corresponding disagreement coefficients $\theta_1, \ldots, \theta_k$. respectively. We can show that the upper bound for rule R3 becomes $\tilde{O}\big(\sqrt[k]{d_1\theta_1 \cdots d_k\theta_k/\varepsilon}\big)$. Furthermore we can find a corresponding lower bound that differs from the upper bound by a factor of $1/k$ only. For the other rules, R1 and R2, we have similar results (that we plan to publish in a follow-up paper). We conjecture that the number of resolution rules, needed to get tight bounds, grows with k. How to prove matching lower and upper bounds for general values of k (up to logarithmic factors, say) is an open problem.

The case of subclasses with infinitely many singletons. The lower bound given in Theorem 6 is only valid for finite s_b^+, s_b^- because the constraint on ε is essentially $1/\varepsilon \geq \max\{s_1^+ s_2^+, s_1^- s_2^-\}$. But even when these parameters are infinite, one easily obtains (from a close inspection of the proof of Theorem 6) a lower bound of $\Omega(1/\varepsilon)$. For special classes, we obtain even larger lower bounds. E.g., let S_∞ denote a class consisting of infinitely many singletons. Then, for the classes $\mathcal{C}_1 = S_\infty^{[d_1]}$ and $\mathcal{C}_2 = S_\infty^{[d_2]}$, a sample of size $\tilde{\Theta}(\min\{d_1, d_2\}/\varepsilon)$ is necessary and sufficient for learning under the Co-training assumption. On the other hand, we get a significantly smaller sample complexity of $\tilde{\Theta}(1/\varepsilon + d/\sqrt{\varepsilon})$ for the classes $\mathcal{C}_1 = \mathcal{C}_2 = S_\infty \cup \{0,1\}^{[d]}$. This shows that we cannot expect to get tight bounds in terms of d and ε alone. To determine how much the Co-training assumption can help for classes with infinite values of s_b^+, s_b^- is work in progress.

Connections to supervised and active learning. 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, or whether existing upper bounds in active learning can be reformulated in completely combinatorial terms using Theorem 4. At least for the realizable case we obtained first results in this direction: we can show that $\tilde{O}(d_\mathcal{C}/\varepsilon)$ unlabeled and $\tilde{O}(\theta \cdot d_\mathcal{H})$ labeled examples are sufficient to learn a concept class of VC-dimension $d_\mathcal{C}$ actively using a hypothesis class of VC-dimension $d_\mathcal{H}$. This is similar to results from [12], but note that our definition of a disagreement coefficient differs from the original definition by Hanneke and, in general, leads to smaller values. For finite values of s_b^+, s_b^-, it is furthermore possible, to replace $\theta \cdot d_\mathcal{H}$ by $s(\mathcal{C})$ and to prove a matching lower bound. We are looking forward to publish the details of these proofs together with further results in a separate paper.

Learning under noise and agnostic learning. The results presented in this paper are only valid for the realizable case of PAC-learning, where we assume that the sample is no subject to noise and that the concept class contains a perfectly compatible target concept. There are several well known generalizations on the PAC framework that incorporate these extensions (e.g. classification noise, malicious noise and agnostic learning). It is an open question if and how our results can be applied to these more relaxed settings.

Acknowledgements. Many thanks to two unknown reviewers for valuable suggestions.

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

[5] 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.

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] Steve Hanneke. Theoretical Foundations of Active Learning. PhD thesis, Carnegie Mellon University, PA, USA, 2009.

[13] Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Proceedings of the 18th Annual Conference on Computational Learning Theory*, pages 127–142, 2005.

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

[15] Aarti Singh, Robert D. Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. In *NIPS*, pages 1513-1520, 2008.

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

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

Appendix A. Table for Theorem 7

Assume $n, k \geq 2$.

	$\Omega\left(\sqrt{\frac{1}{\epsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right)$	$\Omega\left(\sqrt{\frac{1}{\epsilon} \cdot \theta_{max}}\right)$	$\Omega\left(\sqrt{\frac{1}{\epsilon} \cdot \frac{1}{p_{min}}}\right)$
$\mathcal{C}_1$	SF_{kn+2}	SF_{kn+2}	SF_{n+2}
$\mathcal{C}_2$	$\mathrm{co}(\mathrm{SF}_{n+2})$	$\mathrm{co}(\mathrm{SF}_{n+2})$	SF_{n+2}
$p_{min} \in$	$\left[\frac{1}{k}, \frac{1}{2}\right]$	$\left[\frac{1}{kn+2}, \frac{n+2}{kn+2}\right]$	$\left[\frac{1}{(n+2)^2}, \frac{1}{n+2}\right]$
X_1	$\{a_1, \ldots, a_{kn+2}\}$	$\{a_1, \ldots, a_{kn+2}\}$	$\{a_1, \ldots, a_{n+2}\}$
X_2	$\{b_1, \ldots, b_{n+2}\}$	$\{b_1, \ldots, b_{n+2}\}$	$\{b_1, \ldots, b_{n+2}\}$
h_1^*	$\{a_0, a_s\}$	$\{a_0, a_s\}$	$\{a_0, a_s\}$
h_2^*	$X_2 \setminus \{b_0, b_t\}$	$X_2 \setminus \{b_0, b_t\}$	$\{b_0, b_t\}$
$\mathbb{P}(a_s\|+)$	$2\sqrt{\frac{\epsilon}{np_+}}$	$2\sqrt{\frac{k\epsilon}{n}}$	$\sqrt{\frac{\epsilon}{p_+}}$
b_t	$\mathbb{P}(b_t\|-) = 4\sqrt{\frac{\epsilon p_+}{n}}$	$\mathbb{P}(b_t\|-) = 4\sqrt{\frac{\epsilon}{kn}}$	$\mathbb{P}(b_t\|+) = \sqrt{\frac{\epsilon}{p_+}}$
$\mathbb{P}(a_1\|-)$	$1 - \sqrt{\frac{n\epsilon}{p_+}}$	$1 - \sqrt{kn\epsilon}$	$1 - 8n \cdot \sqrt{\epsilon p_+}$
b_1	$\mathbb{P}(b_1\|+) = 1 - \sqrt{\frac{n\epsilon}{p_+}}$	$\mathbb{P}(b_1\|+) = 1 - \frac{1}{p_+}\sqrt{\frac{n\epsilon}{k}}$	$\mathbb{P}(b_1\|-) = 1 - 8n \cdot \sqrt{\epsilon p_+}$
m_0	$\frac{1}{32}\sqrt{\frac{n}{\epsilon p_+}}$	$\frac{1}{32}\sqrt{\frac{kn}{\epsilon}}$	$\frac{1}{80}\sqrt{\frac{1}{\epsilon p_+}}$
$\mathbb{E}[Z_1]$	$p_- \cdot (1 - \mathbb{P}(a_1\|-)) \cdot m \leq \frac{kn}{32}$	$p_- \cdot (1 - \mathbb{P}(a_1\|-)) \cdot m \leq \frac{kn}{32}$	$p_- \cdot (1 - \mathbb{P}(a_1\|-)) \cdot m \leq \frac{n}{10}$
by Markov	$p_1 = \mathbb{P}\left(Z_1 > \frac{kn}{2}\right) < \frac{1}{16}$	$p_1 = \mathbb{P}\left(Z_1 > \frac{kn}{2}\right) < \frac{1}{16}$	$p_1 = \mathbb{P}\left(Z_1 > \frac{n}{2}\right) < \frac{1}{5}$
$\mathbb{E}[Z_2]$	$p_+ \cdot (1 - \mathbb{P}(b_1\|+)) \cdot m \leq \frac{n}{32}$	$p_+ \cdot (1 - \mathbb{P}(b_1\|+)) \cdot m \leq \frac{n}{32}$	$p_- \cdot (1 - \mathbb{P}(b_1\|-)) \cdot m \leq \frac{n}{10}$
by Markov	$p_2 = \mathbb{P}\left(Z_1 > \frac{n}{2}\right) < \frac{1}{16}$	$p_2 = \mathbb{P}\left(Z_1 > \frac{n}{2}\right) < \frac{1}{16}$	$p_2 = \mathbb{P}\left(Z_1 > \frac{n}{2}\right) < \frac{1}{5}$
$\mathbb{E}[\#a_s \text{ occurrance}]$	$p_+ \cdot \mathbb{P}(a_s\|+) \cdot m \leq \frac{1}{16}$	$p_+ \cdot \mathbb{P}(a_s\|+) \cdot m \leq \frac{kp_+}{16} \leq \frac{1}{8}$	$p_+ \cdot \mathbb{P}(a_s\|+) \cdot m \leq \frac{1}{80}$
$\mathbb{P}[a_s \text{ occurrs}]$	$p_3 \leq \frac{1}{16}$	$p_3 \leq \frac{1}{8}$	$p_3 \leq \frac{1}{80}$
$\mathbb{E}[\#b_t \text{ occurrance}]$	$p_- \cdot \mathbb{P}(b_t\|-) \cdot m \leq \frac{1}{8}$	$p_- \cdot \mathbb{P}(b_t\|-) \cdot m \leq \frac{1}{8}$	$p_- \cdot \mathbb{P}(b_t\|+) \cdot m \leq \frac{1}{80}$
$\mathbb{P}[b_t \text{ occurrs}]$	$p_4 \leq \frac{1}{8}$	$p_4 \leq \frac{1}{8}$	$p_4 \leq \frac{1}{80}$
$p_1 + p_2 + p_3 + p_4$	≤ 0.5	≤ 0.5	≤ 0.5
$\mathbb{P}(U \cap E_+)$	$p_+ \cdot \mathbb{P}(a_s\|+)\frac{1-\mathbb{P}(b_1\|+)}{2} \geq \epsilon$	$p_+ \cdot \mathbb{P}(a_s\|+)\frac{1-\mathbb{P}(b_1\|+)}{2} \geq \epsilon$	$p_+ \cdot \mathbb{P}(a_s\|+)\mathbb{P}(b_t\|+) \geq \epsilon$
$\mathbb{P}(U \cap E_-)$	$p_- \cdot \frac{1-\mathbb{P}(a_1\|-)}{2}\mathbb{P}(b_t\|-) \geq \epsilon$	$p_- \cdot \frac{1-\mathbb{P}(a_1\|-)}{2}\mathbb{P}(b_t\|-) \geq \epsilon$	$p_- \cdot \frac{1-\mathbb{P}(a_1\|-)}{2}\frac{1-\mathbb{P}(b_1\|-)}{2} \geq \epsilon$