An Investigation on the Power of Unlabeled Data

Dissertation

zur Erlangung des akademischen Grades eines
Doktor der Naturwissenschaften
der Ruhr-Universität Bochum
in der Fakultät für Mathematik
am Lehrstuhl Mathematik & Informatik

vorgelegt von
Malte Darnstädt

unter der Betreuung von
Prof. Dr. Hans Ulrich Simon

Bochum, Januar 2015

BOCHVM · RVHR-VNIVERSITÄT

Malte Darnstädt
An Investigation on the Power of Unlabeled Data
January 2015
Advisor: Prof. Dr. Hans Ulrich Simon

Ruhr-Universität Bochum
Fakultät für Mathematik
Lehrstuhl Mathematik & Informatik
Universitätsstraße 150
44801 Bochum

Abstract

This thesis studies the power of unlabeled data in classification learning and makes three main contributions:

1.) There is a common intuition, formulated as a rigorous mathematical conjecture by Ben-David, Lu and Pál in [BDLP08], that semi-supervised learning and—even stronger—learning under a fixed distribution reduces the sample complexity by a constant factor only when compared to supervised learning.

We verify this intuition for finite classes and (to a lesser extent) for classes of finite VC-dimension, if the factor is allowed to depend on the concept class. We can remove the dependence on the concept class, if the domain is finite and we allow the learner to fail on a (small) subset of the possible domain distributions.

On the other hand, we prove that the conjecture is in fact wrong for classes of infinite VC-dimension, even if we consider the weaker variant comparing semi-supervised with supervised learning.

2.) Balcan and Blum showed in [BB10], that one labeled example is sufficient for semi-supervised learning in the Co-training framework under the Conditional Independence Assumption. To answer the question how much of this reduction is attributable to unlabeled data and how much is due to Co-training alone, we investigate the reduction achievable by supervised Co-training.

To this end we prove (almost) tight sample size bounds proportional to $\sqrt{1/\varepsilon}$, where ε denotes the accuracy parameter. Thus we see a remarkable reduction by supervised Co-training compared to the standard setting (where the required sample size is proportional to $1/\varepsilon$), albeit not nearly as large as the advantage provided by semi-supervised learners. We make extensive use of a suitably defined variant of Hanneke's disagreement coefficient [Han07], which was developed to analyze agnostic active learning.

We prove (almost) matching worst-case bounds with purely combinatorial parameters, that have a strong connection to Geréb-Graus model of "learning from positive examples only" [GG89].

On a side-note, we settle an open question posed by Auer and Ortner [AO07] about learning intersection-closed classes.

3.) We apply semi-supervised and active learning successfully on the real-world problem of breaking "audio CAPTCHAs", a common scheme for securing internet services against automated attackers.

Acknowledgements

First of all, I have to express my gratitude to my advisor, Professor Dr. Hans Ulrich Simon, without whom this thesis would not have been possible.

I am deeply indebted to the people I had the pleasure to work with during the last years. Special thanks go to Balázs Szörényi, for his hospitality in Szeged, and to my colleagues from the Chair of Theoretical Computer Science at the University of Bochum.

Last, but not least, I would like to thank my parents and family for providing their consistent support.

Contents

Introduction

Over the last decades, machine learning has become a thriving field with many practical applications. We will consider a specific machine learning problem in this thesis, called *classification*: the learning algorithm is presented with a sample of data records (for example, digital audio signals), that come together with labels providing a classification of each record (in our example, the audio signals could be categorized into musical genres; note that we will only consider binary labels in the theoretic part of the thesis). Our goal is to find the rule behind this classification or—if a perfect rule is not conceivable or too hard to find—at least an acceptable approximation.

There are many real-world classification problems, where an abundance of unlabeled data records is readily available (in our example, audio signals could be automatically downloaded from audio and video sharing websites), while classifying unlabeled data is expensive, as this usually requires the work of human experts. It is therefore a natural idea to use the huge amount of unlabeled data to our advantage by providing the learner with an additional, unlabeled sample. It is well-known that so-called *active learning*, where the algorithm is allowed to request the correct labels of a part of the unlabeled sample, can reduce the number of labels necessary to find a good classification rule substantially. However, since providing the requested labels is usually expensive, as a human expert has to assist the active learner, it would be appreciable to make use the unlabeled sample directly. Such learning algorithms are called *semi-supervised* (while learning without utilizing unlabeled data is called *(fully) supervised*). If some notion of compatibility between the data distribution and the labels exists, it is often straightforward to implement semi-supervised learning algorithms. For example, imagine that the data clusters into a small number of sets that are easily detectable; if you assume that each cluster is labeled uniformly, only one labeled sample per cluster is sufficient to discover the classification rule. It is not clear that access to unlabeled data alone provides any advantage if no compatibility assumption holds—in fact, there is a common belief that the benefit of an unlabeled sample is insignificant in that case.

While there are many publications that evaluate semi-supervised learning by conducting experiments on artificial and real-world data sets, the theoretical understanding is still far from complete. The main topic of this thesis is to study the question whether semi-supervised learning *provably helps* in classification tasks—and if it does, to determine the magnitude of its benefit—both with and without compatibility assumptions. Note that we will ignore computational issues (with an exception in the final chapter) and concentrate on information theoretic results.

A second, minor topic is the question, whether unlabeled data helps in real-world classification tasks, one of which we will present in the final chapter.

Organization of the Thesis

This thesis is divided into four parts:

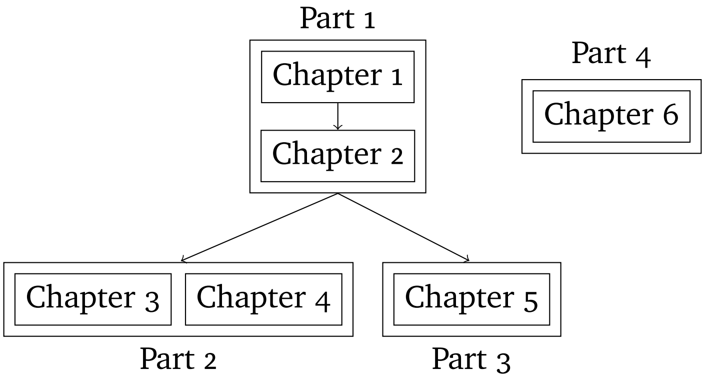
The first part, consisting of Chapters 1 and 2, provides an introduction to learning theory—an analysis of machine learning from the point of view of mathematics and theoretical computer science—and sets the stage for the problems studied in the second and third part. Chapter 1 defines notation and contains introductions to PAC-learning and game theory (these parts can easily be skipped by readers familiar with these topics). Chapter 2 gives an overview over related work and relevant results from learning theory.

In the second part, which consists of Chapters 3 and 4, we investigate the power of unlabeled data in semi-supervised learning when no compatibility between the data distribution and the labels is assumed. We will prove both negative and positive results: we will show in Chapter 3, that semi-supervised learning provides no significant advantage over fully supervised learning in this setting for finite concept classes and (to a lesser extent) for classes of finite VC-dimensions, while we will prove in Chapter 4 that the opposite is true for classes of infinite VC-dimension. The latter result is remarkable, since it defies common intuition about the usefulness of unlabeled data.

In the third part, Chapter 5, we will conduct a case study on a specific compatibility assumption, namely Co-training under the Conditional Independence Assumption. We will see that many labels can be saved by the sole fact that the extra assumption holds, although this saving is not as large as the reduction realized by a semi-supervised learner with access to unlabeled data. One a side-note, we will settle a longstanding open question about learning intersection-closed classes.

The fourth part, which can be read independently from the other parts of the thesis, consists of Chapter 6 and provides an example of a practical application of unlabeled data in machine learning. We will apply both semi-supervised and active learning to successfully ease the breaking of so-called “audio CAPTCHAs”, which are a popular scheme to protect internet services from misuse by automated attackers.

The following illustration gives an overview of the recommended possible reading orders:



Note that this thesis is based on the following peer-reviewed and previously published works, which also contain most results presented here (you will find more detailed explanations, including an overview of previously unpublished results, at the beginning of the relevant chapters):

- "Smart PAC-Learners" [DS11] by the author and Hans Ulrich Simon
- "Unlabeled Data Does Provably Help" [DSS13] by the author, Hans Ulrich Simon and Balázs Szörényi
- "Supervised Learning and Co-training" [DSS14] by the author, Hans Ulrich Simon and Balázs Szörényi
- "The Optimal PAC Bound for Intersection-closed Concept Classes" [Dar15] by the author
- "Reducing the Cost of Breaking Audio CAPTCHAs by Active and Semi-supervised Learning" [DMK14] by the author, Hendrik Meutzner and Dorothea Kolossa

Chapter 1

Basic Definitions and Useful Results

1.1 Notation

First, we define some notation that is used throughout the thesis. Also note the "List of Notations" on page 125.

Vectors and matrices

Vectors are written as bold lowercase letters, e.g., $\boldsymbol{x}, \boldsymbol{b}$, while matrices are written as bold capitals, e.g., $\boldsymbol{A}, \boldsymbol{M}$. The i-th component of a vector $\boldsymbol{x}$ is usually written as x_i and the entry in the i-th row and the j-th column of matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}[i, j]$

For any $f : X \to Y$ and any vector $\boldsymbol{x} = (x_1, \ldots, x_m) \in X^m$, with $m > 0$, we define $f(\boldsymbol{x}) := (f(x_1), \ldots, f(x_m)) \in Y^m$.

Probabilities and random variables

We denote the probability of a (measurable) set A by $\Pr(A)$. If we consider a specific distribution P we often write $P(A)$. The m-fold product of the distribution P is denoted by P^m.

If x is a random variable distributed according to P, we write $x \sim P$. For any event $A(x)$, depending on x, we write the probability that $A(x)$ occurs as $\Pr_{x\sim P}(A(x))$ or, shorter, as $P(A(x))$. For any function f, we denote the expected value of $f(x)$ by $\mathbb{E}_{x\sim P}[f(x)]$. If the random variable and its distribution are clear from context, we often drop the subscript.

Big O notation

We assume that the reader is familiar with Landau's big O notation. The notations $f = \tilde{O}(g)$ and $g = \tilde{\Omega}(f)$ are defined as Landau's O and Ω, but additionally hide logarithmic factors.

1.2 The PAC-Learning Framework

This section provides a succinct introduction into Valiant's PAC-learning framework [Val84] (for a more detailed look at PAC-learning we refer the interested reader to [KV94]). Readers familiar with this topic can easily skip this section.

Preliminaries

Let X denote a non-empty set, called the *domain*, and let P be any distribution over X. We refer to the elements of X as *instances*. Any set $\mathcal{C} \subseteq 2^X$ is called a *concept class* and its elements are called *concepts*. We often identify a concept $c \in \mathcal{C}$ with its indicator function, so that $c : X \to \{0, 1\}$. Let $c^* \in \mathcal{C}$ be some concept, called the *target concept*. We refer to any $\mathcal{H} \subseteq 2^X$, such that $\mathcal{C} \subseteq \mathcal{H}$, as a *hypothesis class* for $\mathcal{C}$. The elements of $\mathcal{H}$ are called *hypotheses*. The *error rate* of a hypothesis $h \in \mathcal{H}$ is given by $P(h \neq c^*) := \Pr_{x\sim P}(h(x) \neq c^*(x))$.[1]

Sample

A *sample* is a pair $(\boldsymbol{x}, \boldsymbol{b})$, such that $\boldsymbol{x} \in X^m$ with $m > 0$ and $\boldsymbol{b} := c^*(\boldsymbol{x})$. (x_i, b_i) is called an *example*, the x_i's are called *sample points* and the b_i's are called *labels*. We refer to an example (x_i, b_i) as *negative* if $b_i = 0$ holds and as *positive* otherwise. For this reason, we sometimes denote negative labels by "$-$" and positive ones by "$+$".

PAC-learner

We say that a (deterministic) algorithm L, often simply called "learner", *PAC-learns concept class $\mathcal{C}$ with hypotheses from $\mathcal{H}$* if the following properties hold:

- The learner's input consists of a random sample $(\boldsymbol{x}, \boldsymbol{b})$, with $\boldsymbol{x} \sim P^m$ and $\boldsymbol{b} = c^*(\boldsymbol{x})$, and two parameters $\varepsilon, \delta > 0$.
- The output is a hypothesis $h \in \mathcal{H}$.
- There exists a function $m_0(\varepsilon, \delta)$, that is polynomial in $1/\varepsilon$ and $\ln(1/\delta)$, such that for any distribution P, any $c^* \in \mathcal{C}$, any $\varepsilon, \delta > 0$ and any $m \geq m_0(\varepsilon, \delta)$ it holds that

$$\Pr_{\boldsymbol{x}\sim P^m} \Big(P(h \neq c^*) \leq \varepsilon \Big) \geq 1 - \delta \ , \tag{1.1}$$

  where h is the output of the learner.

[1] We shortly remark that the existence of all relevant probabilities is guaranteed for "well behaved classes" (see [BEHW89]). Without going into the details, we will always assume that our classes are well behaved.

We call ε the *accuracy* and δ the *confidence* parameter. A hypothesis with an error smaller than ε is called *ε-accurate* and we say that the learner fails, if her hypothesis is not ε-accurate. Note that the learner is not required to observe that her hypothesis is ε-accurate. Inequality (1.1) indicates, that the learner is able to produce an arbitrarily accurate hypothesis with a arbitrary high confidence (the hypothesis is therefore "**p**robably **a**pproximately **c**orrect", i.e., PAC).

Sample complexity

Let us define the *sample complexity of L and $\mathcal{C}$*, denoted by $m_{\mathcal{C}}^{L}$, as the smallest possible function for m_0. Since the (ε, δ)-criterion from (1.1) must hold for *all* pairs of distributions and target concepts, $m_{\mathcal{C}}^{L}$ is valid in the worst case.

We are often interested in the *sample complexity of $\mathcal{C}$*, denoted by $m_{\mathcal{C}}^{*}$, which is given by taking the infimum over all PAC-learners (using arbitrary hypothesis classes). Thus $m_{\mathcal{C}}^{*}$ can be thought of as the worst-case sample complexity of the best possible PAC-learner for $\mathcal{C}$.

Note, that in the original PAC-framework as defined by Valiant, the definition of a PAC-learner demands more than just a polynomial sample complexity. Furthermore, the running time of L must also be polynomial in $1/\varepsilon$ and $\ln(1/\delta)$ (and usually some complexity parameters, see [KV94] for an introduction into the computational details of PAC-learning). Since we do not consider computational or efficiency questions in this thesis (with a minor exception in the final chapter) and are only interested in the size of the sample complexity, we can ignore that demand. Please also note, that all hardness results proven in our information theoretic model are still valid in the computational one.

Consistency and the version space

A hypothesis h that agrees with a sample $(\boldsymbol{x}, \boldsymbol{b})$, i.e., that fulfills $h(\boldsymbol{x}) = \boldsymbol{b}$, is called *consistent* with the sample. The set of all consistent hypotheses in $\mathcal{H}$ is called the *version space* of $(\boldsymbol{x}, \boldsymbol{b})$:

$$V_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{b}) := \{h \in \mathcal{H} : h(\boldsymbol{x}) = \boldsymbol{b}\}$$

For convenience we define the following notation for $c^* \in \mathcal{C}$ and $X' \subseteq X$:

$$\begin{aligned} V_{\mathcal{H}}(\boldsymbol{x}, c^*) &:= V_{\mathcal{H}}(\boldsymbol{x}, c^*(\boldsymbol{x})) \\ V_{\mathcal{H}}(X', c^*) &:= \{h \in \mathcal{H} : h(x) = c^*(x) \text{ for all } x \in X'\} \end{aligned}$$

The Vapnik-Chervonenkis dimension

An important combinatorial parameter for proving lower and upper bounds on the sample complexity was defined by Vapnik and Chervonenkis [VC71]:

we say that a set $X' \subseteq X$ is *shattered* by $\mathcal{C}$ if for any $X'' \subseteq X'$ there is a concept $c \in \mathcal{C}$ such that $c \cap X' = X''$ (so that concepts from $\mathcal{C}$ can induce arbitrary labelings on X'). The *VC-dimension* of $\mathcal{C}$, denoted by $\mathrm{VCdim}(\mathcal{C})$, is defined as the size of the largest set shattered by $\mathcal{C}$ or as infinite if no such set exists.

The covering number

If the distribution P is fixed, the following parameter can provide stronger bounds on the sample complexity: a subset $\mathcal{C}' \subseteq \mathcal{C}$ is called an *ε-cover* of $\mathcal{C}$ with respect to P if for any $c \in \mathcal{C}$ there exists a $c' \in \mathcal{C}$ with $P(c \neq c') \leq \varepsilon$. The *covering number*, denoted by $N_{\mathcal{C},P}(\varepsilon)$, is defined as the size of the smallest ε-cover of $\mathcal{C}$ with respect to P (which can be infinite).

Variants of PAC-learners

Besides the basic definition, we consider the following special cases and variants of PAC-learners:

Consistent learners A learner who always outputs a consistent hypothesis is called a *consistent learner*.

Proper learners A learner is called *proper*, if she always outputs a hypothesis from $\mathcal{C}$, i.e., if $\mathcal{H} = \mathcal{C}$ holds. Otherwise the learner is called *improper*.

Randomized learners If we allow learners to be randomized algorithms, the outer probability in (1.1) must also be taken over the internal randomization of the learner.

A learner that makes no use of randomization is called *deterministic*.

Semi-supervised learning In semi-supervised learning the learner has an additional finite *unlabeled* sample at her disposal. Like in the labeled sample, the unlabeled sample points are drawn independently and distributed according to P.

A learner that has no additional information about the distribution P is called *fully supervised* or simply *supervised*.

Note that the size of the unlabeled sample is not added to the sample complexity, which is defined as the number of necessary *labeled* examples. Furthermore, the sample complexity of a class $\mathcal{C}$ in the semi-supervised setting can never be larger than the sample complexity in the fully supervised setting, because a semi-supervised learner can choose to ignore her unlabeled sample and behave just like a supervised learner.

Learning under a fixed distribution The domain distribution P is fixed and the learning algorithm may be thought of being tailored to P or receiving P as part of her input.

Note that we can regard a learner tailored to P as a "supercharged" semi-supervised learner equipped with an unlabeled sample so large that she knows P perfectly well. Moreover, the sample complexity of a class $\mathcal{C}$ in the fixed distribution setting can never be larger than the sample complexity in the semi-supervised setting, because a learner who knows P can always simulate a semi-supervised learner by generating an unlabeled sample.

We denote the *sample complexity of L, $\mathcal{C}$ and P* in this framework by $m^L_{\mathcal{C},P}$. If P is more benign than the worst-case distribution, this value can be much smaller than $m^L_{\mathcal{C}}$. We will investigate the question, whether this advantage can only be realized by learners L that are tailored to distribution P.

Similar as before we define the *sample complexity of $\mathcal{C}$ and P*, denoted by $m^*_{\mathcal{C},P}$, by taking the infimum over all PAC-learners. Note that this value can again be much smaller than $m^*_{\mathcal{C}}$.

Active learning In active learning the learner has access to an additional unlabeled sample, similar as in semi-supervised learning, but here she is allowed to ask an oracle for the labels of selected sample points in her unlabeled sample.

A non-active learner is called *passive*.

Agnostic learning and learning under noise We assumed the existence of a target concept $c^* \in \mathcal{C}$, that can explain the labels of all sample points perfectly well, and that the labels in the sample are undisturbed by noise. Of course, these requirements are often not satisfied in practical applications of machine learning.

There are several well-known variants of PAC-learning that relax these restrictions; for example *learning under noise*, where each label is wrong with a certain probability, and *agnostic learning*, where we do not assume the existence of a target concept and the examples are generated by a distribution over $X \times \{0, 1\}$ instead. In agnostic learning we demand that the learner finds a hypothesis that is ε-close to the best hypothesis in $\mathcal{C}$, which may have a non-zero error.

The non-agnostic setting without noise, in which will work in this thesis, is called the *realizable setting*.

Now we can characterize our original definition of PAC-learning with this vocabulary as deterministic, fully supervised, passive and realizable. In the case $\mathcal{C} \neq \mathcal{H}$, which we allowed, the learner is also improper.

A game for proving lower bounds

To prove lower bounds on the sample complexity we often analyze the size of the *Bayes-error*, which is the smallest possible error rate achievable by the optimal learner, who proceeds according to the so-called *Bayes-decision*.

Furthermore, we usually consider the following game between the (deterministic) learner and an "adversary":

The adversary chooses a probability distribution D on pairs (c^*, P) where $c^* \in \mathcal{C}$ and P is a probability distribution over X. To create a sample we first draw a target concept c^* and a domain distribution P according to D, then we generate sample points $\boldsymbol{x}$ according to P^m and compute the corresponding labels $\boldsymbol{b} = c^*(\boldsymbol{x})$. Because proofs can become simpler if the learner is given additional information about the learning task (e.g., it may be easier to determine the Bayes-decision), we may add auxiliary information to the learner's input. As usual, the learner outputs a hypothesis h. She fails if h is not ε-accurate, i.e., if $P(h \neq c^*) > \varepsilon$ holds.

There are two main differences between this game and PAC-learning:

- We are switching from PAC-learning's worst-case analysis, where the learner has to cope with any pair (c^*, P), to an average case analysis.
- The additional information given to the learner is missing in PAC-learning.

Because both changes can only be advantageous to the learner, they pose no problems for the prove of lower bounds: if the learner fails with a probability larger than δ, she will also miss the (ε, δ)-criterion of PAC-learning (as given in (1.1)). Furthermore, Yao's principle [Yao77] states that lower bounds obtained by this technique are even valid for randomized learners.

1.3 Game Theory and the Minimax Theorem

In this section we give a short introduction to the theory of two-player zero-sum games and present the Minimax Theorem (for a more detailed introduction we refer the interested reader to [Mor94]). Readers familiar with game theory can easily skip this section.

Two-player zero-sum games and payoff matrices

We want to model games with two players, called Alice and Bob. We assume that the game is deterministic and that all information about the game is open for both to see (there are no dice or hidden cards, for example). Both players play a single move simultaneously before the outcome of the game is determined. Let us say that Alice can choose between m different moves, while Bob has n different moves his disposal.

When Alice plays move i and Bob plays move j, we denote Bob's outcome of the game by a real number $\boldsymbol{A}[i,j]$. A higher value corresponds to a better outcome for Bob. An analogue outcome $\boldsymbol{A}'[i,j]$ can be defined for Alice. A game is called *zero-sum*, if Bob wins exactly the amount that Alice loses and vice versa, i.e., if $\boldsymbol{A}[i,j] = -\boldsymbol{A}'[i,j]$ holds for all i, j. The matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is called the *payoff matrix* of the game.

Mixed and pure strategies

We allow Alice and Bob to choose their moves randomly; thus Alice's strategy is given by a probability vector $\boldsymbol{p} \in [0,1]^m$ (with $\sum_{i=1}^{m} p_i = 1$), such that the probability of playing move i is exactly p_i. Analogously, Bob's strategy is given by a probability vector $\boldsymbol{q} \in [0,1]^n$. We refer to these randomized strategies as *mixed strategies*, while strategies that pick one fixed move are called *pure*. The expected gain of Bob (and the expected loss of Alice) is now given by

$$\boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} \ . \tag{1.2}$$

Thus Alice's goal is to the minimize the value from (1.2), while Bob tries to maximize it.

Optimal strategies and the Minimax Theorem

We now assume that one player, say Alice, makes the first draw by fixing her mixed strategy $\boldsymbol{p}$. Now Bob can easily optimize his strategy $\boldsymbol{q}$ based on the choice of Alice (in fact, Bob's optimal strategy is a pure one: he can play move j, where j is the index of the largest component of $\boldsymbol{p}^\top \boldsymbol{A}$). Thus Alice's optimal choice in this situation is playing a mixed strategy $\boldsymbol{p}$ that minimizes the largest component of $\boldsymbol{p}^\top \boldsymbol{A}$. An analogue remark applies if Bob makes the first draw.

Now John von Neumann's famous Minimax Theorem states the following:

Theorem 1 (Minimax [vN28]). *For all $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ holds*

$$\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} = \max_{\boldsymbol{q}} \min_{\boldsymbol{p}} \boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} \ , \tag{1.3}$$

where the minium and maximum is taken over probability vectors (i.e., $\boldsymbol{p} \in [0,1]^m$, $\boldsymbol{q} \in [0,1]^n$ and $\sum_{i=0}^{n} p_i = \sum_{j=0}^{m} q_j = 1$).

A proof can be found in [Mor94].

Note that the left- resp. right-hand side of (1.3) models the situation where Alice resp. Bob makes the first draw. Thus, the Minimax Theorem states that it is irrelevant who fixes their strategy first, as long as both players play an optimal strategy. Moreover, it follows that we can assign a unique real *value* to each game, namely the one from equation (1.3).

1.4 A Useful Result From Probability Theory

We assume that the reader is familiar with basic probability theory including results like the Union Bound and Markov Inequality. In this section, we present the following "concentration bound", which we will apply several times in the later chapters:

Theorem 2 (Chernoff-bounds [Che52]). *Let $X_1, \ldots, X_m$ be a sequence of independent, identically distributed Bernoulli random variables with $P(X_i = 1) = p$ for all $1 \leq i \leq m$. Then for all $0 \leq \gamma \leq 1$ the following holds:*

$$\Pr\Big(\sum_{i=1}^{m} X_i > (1+\gamma)\cdot pm\Big) \leq e^{-mp\gamma^2/3}$$

and

$$\Pr\Big(\sum_{i=1}^{m} X_i < (1-\gamma)\cdot pm\Big) \leq e^{-mp\gamma^2/2}$$

Thus the Chernoff-bounds state that a binomially distributed random variable is close to its expected value with a high probability. The theorem as it is given here is a special case of a more general result, whose proof can be found in Chernoff's original paper [Che52].

Chapter 2

Related Work From Learning Theory

This chapter serves two purposes: we will survey related work from learning theory and, at the same time, present several results about the sample complexity—many of them classic theorems in the PAC-learning framework—which we will apply in later chapters. Note that Chapters 5 and 6 contain additional sections discussing related work, as these chapters have further connections to fields not mentioned here.

This chapter contains only one previously unpublished result (see Lemma 8).

2.1 Work on Supervised Learning

The following classic results give almost tight bounds on the sample complexity of supervised PAC-learning in the worst case:

Theorem 3 ([BEHW87]). *Let $\mathcal{C} \subseteq \mathcal{H}$ be finite concept and hypothesis classes. Then $\mathcal{C}$ can be PAC-learned by any algorithm returning a consistent hypothesis from $\mathcal{H}$ using*

$$m = \left\lceil \frac{\ln |\mathcal{H}| + \ln(1/\delta)}{\varepsilon} \right\rceil$$

labeled examples. With the choice $\mathcal{H} = \mathcal{C}$ the learner is proper and yields the same upper bound on the sample complexity, i.e., $m^*_{\mathcal{C}}(\varepsilon, \delta) = O((\ln |\mathcal{C}| + \ln(1/\delta))/\varepsilon)$.

Theorem 4 ([BEHW89]). *Let $\mathcal{C} \subseteq \mathcal{H}$ be a concept and hypothesis class of finite VC-dimensions. Then $\mathcal{C}$ can be PAC-learned by any algorithm returning a consistent hypothesis from $\mathcal{H}$ using*

$$m = O\left(\frac{\mathrm{VCdim}(\mathcal{H}) \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon} \right)$$

labeled examples. With the choice $\mathcal{H} = \mathcal{C}$, the learner is proper and yields the same upper bound on the sample complexity, i.e., $m^*_{\mathcal{C}}(\varepsilon, \delta) = O((\mathrm{VCdim}(\mathcal{C}) \ln(1/\varepsilon) + \ln(1/\delta))/\varepsilon)$.

Theorem 5 ([EHKV89]). *Any algorithm PAC-learning a class $\mathcal{C}$ of finite VC-dimension needs at least*

$$m^*_{\mathcal{C}}(\varepsilon, \delta) = \Omega\left(\frac{\mathrm{VCdim}(\mathcal{C}) + \ln(1/\delta)}{\varepsilon}\right)$$

many labeled examples in the worst case (for small enough $\varepsilon, \delta > 0$).

Note that the lower bound matches with the upper bound from Theorem 4 up to a logarithmic factor. It is still an open question, whether this factor of $\ln(1/\varepsilon)$ can be overcome; see Section 5.3.5 for a positive answer in the case of concept classes that are closed under intersection.

2.2 Work on Learning with Unlabeled Data

As mentioned in the introduciton, obtaining labeled samples is expensive in many applications, since it often requires the work of a human expert, while acquiring large amounts of unlabeled data is relatively cheap.

This motivates the question if it is possible to reduce the sample complexity $m^*_{\mathcal{C}}$ by exploiting a large amount of unlabeled data, which is the main topic of this thesis.

Active learning

This question is answered in the positive in the framework of active learning, where the learner is allowed to ask an oracle for the true label of a part of her unlabeled data sample. Since active learning is relatively well studied and understood (even in the agnostic setting; see [CAL94, Das05, BBZ07, DHM07, Han07, BBL09, BHV10, BL13]), we will turn our focus to semi-supervised learning[1], where the learner has access to an additional unlabeled sample but stays passive, i.e., she may not ask an oracle for labels. We will also investigate the setting of "learning under a fixed distribution" as an idealized and often easier to analyze version of semi-supervised learning.

Utilizing unlabeled data with extra assumptions

In the framework of semi-supervised learning one usually makes use of *extra assumptions*, that provide some kind of compatibility between the distribution P and the target concept c^*. You will find a discussion of the most popular extra

[1] While there are many publications on semi-supervised learning, the majority of them only considers practical applications. We will contribute our share in Chapter 6, where we apply semi-supervised and active learning to a real-world problem from the field of internet security.

assumptions in the introduction of [CSZ06]. It is known that semi-supervised learning can reduce the need of labeled samples enormously, if a suited extra assumption holds. For example, Balcan and Blum proved in [BB10] that only *one* labeled example (and polynomially many unlabeled ones) suffices for PAC-learning in a framework called "Co-training" if the "Conditional Independence Assumption" holds. We will take an intensive look at this setting in Chapter 5. For an analysis of semi-supervised learning in an augmented version of PAC-learning, that is able to model many popular extra assumptions, we refer to [BB10].

Utilizing unlabeled data without extra assumptions

As stated in the introduction, there is a common intuition that the effect of unlabeled data is marginal if extra assumptions are absent.

Note that the lower bound from supervised PAC-learning, Theorem 5, is still valid if the learner is allowed to know the domain distribution. So we already see that no amount of unlabeled data can help to reduce the sample complexity *in the worst case*. On the other hand, Theorem 5 does not rule out the possibility that unlabeled data can still be helpful if the domain distribution is more benign.

Already in 1991 Benedek and Itai proved the following upper bound on the sample complexity in the fixed-distribution setting:

Theorem 6 ([BI91]). *Let $\mathcal{C} \subseteq \mathcal{H}$ be a concept and hypothesis class with a finite covering number $N_{\mathcal{H},P}(\varepsilon)$ for all $\varepsilon > 0$. Then $\mathcal{C}$ can be PAC-learned under a fixed distribution P using*

$$m = O\left(\frac{\ln(N_{\mathcal{H},P}(\varepsilon/2)) + \ln(1/\delta)}{\varepsilon}\right)$$

labeled examples. With the choice $\mathcal{H} = \mathcal{C}$, the learner is proper and yields the same upper bound on the sample complexity under the fixed distribution P, i.e., $m^*_{\mathcal{C},P}(\varepsilon,\delta) = O(\ln(N_{\mathcal{C},P}(\varepsilon/2) + \ln(1/\delta))/\varepsilon)$.

In the same paper, Benedek and Itai also showed a general lower bound on the sample size in the fixed distribution setting (based on packing numbers). A possible way to verify the "intuition" mentioned above would be to show that a fully supervised PAC-learner exists whose sample complexity is in the same order as the lower bound from Benedek and Itai. Since their lower bound does not even match the upper bound from Theorem 6, this looks like a hopeless approach.

In 2005, Kääriäinen [Kää05] developed a semi-supervised learning strategy, which does not rely on extra assumptions and is able to reduce the sample complexity by a factor of up to two. The basic idea is to search for the *center* of the version space (measured by a pseudometric given by P) instead of returning some arbitrary consistent hypothesis. Since a reduction of the sample complexity by a constant factor is asymptotically negligible, Kääriäinen's result is in accordance with the common intuition.

In 2010 Balcan and Blum showed[2] that the upper bound from Theorem 6 can be transferred to the semi-supervised setting (without relying on extra assumptions):

Theorem 7 ([BB10]). *Let $\mathcal{C} \subseteq \mathcal{H}$ be a concept and hypothesis class of finite VC-dimensions and finite covering numbers. Then $\mathcal{C}$ can be PAC-learned in the semi-supervised setting from*

$$m = O\left(\frac{\ln(N_{\mathcal{H},P}(\varepsilon/6)) + \ln(1/\delta)}{\varepsilon}\right)$$

labeled and

$$m_U = O\left(\frac{\mathrm{VCdim}(\mathcal{H})\ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon^2}\right)$$

unlabeled examples. With the choice $\mathcal{H} = \mathcal{C}$, the learner is proper and yields the same upper bound on the sample complexity in semi-supervised learning, i.e., there exists a semi-supervised learner L such that for all domain distributions P holds $m^L_{\mathcal{C},P}(\varepsilon,\delta) = O(\ln(N_{\mathcal{C},P}(\varepsilon/6) + \ln(1/\delta))/\varepsilon)$.

The conjecture of Ben-David, Lu and Pál

Ben-David, Lu and Pál formulated the "common intuition" as a rigorous mathematical statement. They conjectured in [BDLP08] that even knowing P perfectly can reduce the sample complexity by a constant factor only, i.e., for any class $\mathcal{C}$ there exists a fully supervised learner L and a constant $k \geq 1$ such that for any distribution P holds

$$m^L_{\mathcal{C},P}(\varepsilon,\delta) \leq k \cdot m^*_{\mathcal{C},P}(\varepsilon,\delta) \tag{2.1}$$

for small enough $\varepsilon, \delta > 0$.

Ben-David et al. could prove this conjecture for several basic classes over the real line under continuous domain distributions (with $k \approx 2$). Nevertheless, the following lemma observes that (2.1) is trivially wrong. However, this is only the case in letter but not in spirit, as a slight change in the formulation of (2.1) yields a similar conjecture that is not so easy to falsify.

[2] To see the result, apply Theorem 13 from [BB10] with a constant compatibility of 1 for all concepts and distributions.

Lemma 8. *There is a concept class $\mathcal{C}$ such that for all $0 < \varepsilon < 1/2$ there exists a family of distributions $\mathcal{P}$ such that the following holds:*

1. *For every $P \in \mathcal{P}$ exists a learner (tailored to P) that always finds an ε-accurate hypothesis without seeing any labeled examples, i.e., for all $P \in \mathcal{P}$ and all non-negative δ holds $m^*_{\mathcal{C},P}(\varepsilon, \delta) = 0$.*
2. *For any fully supervised learner L and all $\delta < 1/36$ exists a distribution $P \in \mathcal{P}$ such that $m^L_{\mathcal{C},P}(\varepsilon, \delta) > 0$.*

Proof. Let $X = \{0, \dots, 6\}$ and $\mathcal{C} = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \{1, \dots, 6\}\}$. $\mathcal{P}$ consists of the following distributions: for any $i \in \{1, 2, 3\}$ and $j \in \{4, 5, 6\}$ let $P_{i,j}$ be the distribution that assigns probability $1 - 2\varepsilon$ to the point 0 and evenly distributes the remaining probability mass of 2ε over the four points $X \setminus \{0, i, j\}$ (the points i and j have zero probability mass).

To prove the first part of the lemma, let $i' \in \{1, 2, 3\} \setminus \{i\}$ and $j' \in \{4, 5, 6\} \setminus \{j\}$. Observe that the hypothesis $h = \{i', j'\}$ is ε-accurate for any choice of $c^* \in \mathcal{C}$ under distribution $P_{i,j}$. Therefore, a learner that knows $P_{i,j}$ can simply output h without seeing any labeled examples.

To prove the lower bound we consider a game between the learner and an adversary. The adversary chooses a distribution $P_{i,j}$ and a target $c^* \in \mathcal{C}$ uniformly at random. It suffices to show that any fixed hypothesis $h \subseteq X$ fails to achieve ε-accuracy with a probability of at least $1/36$. We can easily show the existence of a "bad pair" $(c^*, P_{i,j})$ for any fixed h:

- If h contains no points from $\{1, 2, 3\}$, we include $\{1, 2, 3\}$ in c^* and choose i arbitrarily.
- If h contains one point i' from $\{1, 2, 3\}$, we include $\{1, 2, 3\}$ in c^* and let $i = i'$.
- If h contains two points i', i'' from $\{1, 2, 3\}$, we exclude $\{1, 2, 3\}$ from c^* and choose $i \in \{1, 2, 3\}$ such that $i' \neq i \neq i''$.
- If h contains $\{1, 2, 3\}$ completely, we exclude $\{1, 2, 3\}$ from c^* and choose i arbitrarily.

In any case, the error rate of h on $\{1, 2, 3\}$ is at least ε. An analogous remark applies to h on $\{4, 5, 6\}$, so that the overall error rate of h is at least 2ε. This concludes the proof, since the probability of drawing a bad pair $(c^*, P_{i,j})$ is at least $1/|\mathcal{C} \times \mathcal{P}| = 1/36$. □

In light of Lemma 8, we will allow the fully supervised learner in (2.1) to be twice as inaccurate as the learner tailored to the domain distribution, i.e., the conjecture now states: for any class $\mathcal{C}$ there exists a fully supervised learner L and a constant $k \geq 1$ such that for any distribution P holds

$$m^L_{\mathcal{C},P}(2\varepsilon, \delta) \leq k \cdot m^*_{\mathcal{C},P}(\varepsilon, \delta) \tag{2.2}$$

for small enough $\varepsilon, \delta > 0$.

We call the constant k from (2.2) the *performance ratio*. As we will show in the following two chapters, the altered conjecture is true for many concept classes, but not for all.

Chapter 3

Where Unlabeled Data Does Not Help

This chapter is based on the following article resp. conference paper:

- "Smart PAC-Learners" [DS11][1], which is a joint work of the author with Hans Ulrich Simon (please note that this article is partly based on an earlier conference paper by Hans Ulrich Simon [Sim09], which proved the existence of PAC-learners with a universal constant performance ratio over finite domains for special classes with an additional property only)
- "Unlabeled Data Does Provably Help" [DSS13][2], which is a joint work of the author with Hans Ulrich Simon and Balázs Szörényi

All results in this chapter were originally published in these two works.

3.1 Introduction

In this chapter we investigate the conjecture (2.2), that there exists a fully supervised learner whose label consumption exceeds the label consumption of the best learner with full prior knowledge of the domain distribution at most by a factor $k(\mathcal{C})$ that depends on $\mathcal{C}$ only, as opposed to a dependence on ε or δ. Furthermore, we will also examine the possibility to remove the dependence of k on the class $\mathcal{C}$.

3.1.1 Main Results

Finite classes and classes with finite VC-dimension

The following result, whose proof is found in Section 3.2, confirms conjecture (2.2) to a large extent for finite classes, and to a somewhat smaller extent

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

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

for classes of finite VC-dimension:

Theorem 9. *Let $\mathcal{C}$ be a concept class over domain X that contains the constant zero- and the constant one-function. Then:*

1. *If $\mathcal{C}$ is finite, there exists a proper and fully supervised PAC-learning algorithm L such that, for every domain distribution P,*

$$m^L_{\mathcal{C},P}(2\varepsilon,\delta) = O(\ln|\mathcal{C}|)\cdot m^*_{\mathcal{C},P}(\varepsilon,\delta) \ .$$

2. *If the VC-dimension of $\mathcal{C}$ is finite, there exists a proper and fully supervised PAC-learning algorithm L such that, for every domain distribution P,*

$$\begin{aligned} m^L_{\mathcal{C},P}(2\varepsilon,\delta) &= O(\mathrm{VCdim}(\mathcal{C})\cdot\ln(1/\varepsilon))\cdot m^*_{\mathcal{C},P}(\varepsilon,\delta) \\ &= \tilde{O}(\mathrm{VCdim}(\mathcal{C}))\cdot m^*_{\mathcal{C},P}(\varepsilon,\delta) \ . \end{aligned}$$

Learners with universal performance ratios

For classes over *finite* domains X and an inequality slightly different from (2.2), we can show the existence of proper and fully supervised learners, that achieve a performance ratio k, which has no dependence on $\mathcal{C}$. To this end, it will be convenient to think of the concept class $\mathcal{C}$, the sample size m and the accuracy parameter ε as fixed, respectively, and to figure out the smallest value for δ such that m examples suffice to meet the (ε,δ)-criterion of PAC-learning. Let $\delta^*_P(\varepsilon,m)$ denote the smallest value for δ that can be achieved by a learner who is specialized to P and, for any learner L, let $\delta^L_P(\varepsilon,m)$ denote the smallest value of δ such that m examples are sufficient to meet the (ε,δ)-criterion from inequality (1.1) provided that L was exposed to the domain distribution P.

We aim at an inequality of the form

$$\delta^L_P(2\varepsilon,m) \le k\cdot\delta^*_P(\varepsilon,m) \tag{3.1}$$

where k denotes a constant not even depending on $\mathcal{C}$.

Because of this independence, the statement is stronger than the result from Theorem 9. However, the result, that we can actually prove, has one important limitation (besides the finiteness of X), which makes it weaker and therefore incomparable to Theorem 9: inequality (3.1) is not verified for all distributions P but for a volume of approximately $1-2/k$, say when measured with the uniform distribution over the $|X|$-dimensional probability simplex (whose points represent distributions over X). Actually the final result is somewhat more general: we can fix any measure μ on the $|X|$-dimensional probability simplex (expressing which domain distributions we consider as particularly important), and still achieve validity of (3.1) for a volume of approximately $1-2/k$ of all domain distributions. Clearly, the design of learner L changes when μ changes.

The formal statement is found in Theorem 19.

3.2 Proof of Theorem 9

We start with the following lower bound on $m^*_{\mathcal{C},P}(\varepsilon,\delta)$:

Lemma 10. *Let $\mathcal{C}$ be a concept class and let P be a distribution on domain X. For any $\varepsilon > 0$, let*

$$\lceil\varepsilon\rceil_{\mathcal{C},P} = \min\{\varepsilon' \mid (\varepsilon' \geq \varepsilon) \wedge (\exists c, c' \in \mathcal{C} : P(c \neq c') = \varepsilon'\}$$

where, by convention, the minimum of an empty set equals ∞. With this notation, the following holds:

1. *If $\lceil 2\varepsilon\rceil_{\mathcal{C},P} \leq 1$, then $m^*_{\mathcal{C},P}(\varepsilon,\delta) \geq 1$.*
2. *Let $\gamma = 1 - \lceil 2\varepsilon\rceil_{\mathcal{C},P}$. If $\lceil 2\varepsilon\rceil_{\mathcal{C},P} < 1$, then*
$$m^*_{\mathcal{C},P}(\varepsilon,\delta) \geq \log_{1/\gamma}\frac{1}{2\delta} = \Omega\left(\log_{1/\gamma}\frac{1}{\delta}\right) \ . \tag{3.2}$$
3. *If $\lceil 2\varepsilon\rceil_{\mathcal{C},P} \leq 1/4$, then*
$$m^*_{\mathcal{C},P}(\varepsilon,\delta) \geq \left\lfloor \frac{\ln(1/(2\delta))}{2\lceil 2\varepsilon\rceil_{\mathcal{C},P}} \right\rfloor = \Omega\left(\frac{\ln(1/\delta)}{\lceil 2\varepsilon\rceil_{\mathcal{C},P}}\right) \ . \tag{3.3}$$

Proof. It is easy to see that at least one labeled example is needed if $\lceil 2\varepsilon\rceil_{\mathcal{C},P} \leq 1$. Let us now assume that $\lceil 2\varepsilon\rceil_{\mathcal{C},P} < 1$. Let $c, c' \in \mathcal{C}$ be chosen such that $P(c \neq c') = \lceil 2\varepsilon\rceil_{\mathcal{C},P}$. The adversary picks c and c' as target concept with probability $1/2$, respectively. With a probability of $(1 - \lceil 2\varepsilon\rceil_{\mathcal{C},P})^m$, none of the labeled examples hits the symmetric difference of c and c'. Since $P(c \neq c') \geq 2\varepsilon$, the learner has no hypothesis at her disposal that is ε-accurate for c and c'. Thus, if none of the examples distinguishes between c and c', the learner will fail with a probability of $1/2$. We can conclude that the learner fails with an overall probability of at least $\frac{1}{2}(1 - \lceil 2\varepsilon\rceil_{\mathcal{C},P})^m = \frac{1}{2}\gamma^m$.

Setting this probability less than or equal to δ and solving for m leads to the lower bound (3.2). If $\lceil 2\varepsilon\rceil_{\mathcal{C},P} \leq 1/4$, a straightforward computation shows that $\frac{1}{2}\gamma^m$ is bounded from below by $\frac{1}{2}\exp(-2\lceil 2\varepsilon\rceil_{\mathcal{C},P} \cdot m)$. Setting this expression less than or equal to δ and solving for m leads to the lower bound (3.3). □

We are ready now for the proof of Theorem 9:

Proof. We use the notation from Lemma 10. We first present the main argument under the (wrong!) assumption that $\lceil\varepsilon\rceil_{\mathcal{C},P}$ is known to the learner. At the end of the proof, we explain how a fully supervised learning algorithm can compensate for not knowing P. The first important observation, following directly from the definition of $\lceil\varepsilon\rceil_{\mathcal{C},P}$, is that, in order to achieve an accuracy of ε, it suffices to achieve an accuracy $\lceil \varepsilon \rceil_{\mathcal{C},P}$ with a hypothesis from $\mathcal{C}$. Thus, for the purpose of Theorem 9, it suffices to have a supervised proper learner that achieves accuracy $\lceil 2\varepsilon \rceil_{\mathcal{C},P}$ with confidence δ. We proceed with the following case analysis:

Case 1: $\lceil 2\varepsilon \rceil_{\mathcal{C},P} \leq 1/4$.
There is a gap of $O(\ln |\mathcal{C}|)$ only between the upper bound from Theorem 3 (with $\lceil 2\varepsilon \rceil_{\mathcal{C},P}$ in the role of ε) and the lower bound (3.3). Returning a consistent hypothesis, so that Theorem 3 applies, is appropriate in this case.

Case 2: $1/4 < \lceil 2\varepsilon \rceil_{\mathcal{C},P} < 15/16$.
We may argue similarly as in Case 1 except that the upper bound from Theorem 3 is compared to the lower bound (3.2). (Note that $\gamma = \theta(1)$ in this case.) As in Case 1, returning a consistent hypothesis is appropriate.

Case 3: $15/16 < \lceil 2\varepsilon \rceil_{\mathcal{C},P} < 1$.
In this case $0 < \gamma = 1 - \lceil 2\varepsilon \rceil_{\mathcal{C},P} < 1/16$. The learner will exploit the fact that one of the hypotheses $\emptyset$ and X is a good choice. She returns hypothesis X if label "1" has the majority within the labeled examples, and hypothesis $\emptyset$ otherwise. Let c^*, as usual, denote the target concept. If $\gamma < P(c^*) < 1 - \gamma$, then both of $\emptyset$ and X are $\lceil 2\varepsilon \rceil_{\mathcal{C},P}$-accurate. Let us assume that $P(c^*) \leq \gamma$. (The case $P(c^*) \geq 1-\gamma$ is symmetric.) The learner will fail only if, despite of the small probability γ for label "1", these labels have the majority. It is easy to see that the probability for this to happen is bounded by $(m/2)\binom{m}{m/2}\gamma^{m/2}$ and, as seen by an application of Stirling's Formula, also bounded by $2^{3m/2}\gamma^{m/2} = (8\gamma)^{m/2}$.

Setting the last expression less than or equal to δ and solving for m reveals that $O(\log_{1/\gamma}(1/\delta))$ many labeled examples are enough. This matches the lower bound (3.2) modulo a constant factor.

Case 4: $\lceil 2\varepsilon \rceil_{\mathcal{C},P} = 1$.
This is a trivial case where each labeled example almost surely makes inconsistent any hypothesis $h \in \mathcal{C}$ of error at least ε. The learner may return any hypothesis that is supported by at least one labeled example.

Case 5: $\lceil 2\varepsilon \rceil_{\mathcal{C},P} = \infty$.
This is another trivial case where any concept from $\mathcal{C}$ is 2ε-accurate with respect to any other concept from $\mathcal{C}$. The learner needs no labeled example and may return any $h \in \mathcal{C}$.

In any case, the performance ratio is bounded by $O(\ln |\mathcal{C}|)$. We finally have to explain how this can be exploited by a supervised learner L who does not have any prior knowledge of P. The main observation is that, according to the bound in Theorem 3, the condition

$$m > \left\lceil \frac{\ln(|\mathcal{C}|/\delta)}{15/16} \right\rceil \tag{3.4}$$

indicates that the sample size is large enough to achieve an accuracy below $15/16$ so that returning a consistent hypothesis is the appropriate action (as in Cases 1 and 2 above). If, on the other hand, condition (3.4) is violated, then L will set either $h = \emptyset$ or $h = X$ depending on which label holds the majority (which would also be an appropriate choice in Cases 3 and 4 above). It is not hard to show that this procedure leads to the desired performance, which concludes the proof for the first part of Theorem 9.

As for the second part, one can use an analogous argument that employs Theorem 4 instead of Theorem 3. □

3.3 Existence Proof of Learners with Universal Performance Ratios

Throughout this section—as already mentioned in the introduction—we think of $\mathcal{C}$, m and ε as fixed, and we consider the quantities $\delta_P^L(2\varepsilon, m)$ and $\delta_P^*(\varepsilon, m)$ instead of $m_{\mathcal{C},P}^L(2\varepsilon, \delta)$ and $m_{\mathcal{C},P}^*(\varepsilon, \delta)$.

Since X is now a finite domain, let $d := |X|$ and $X = \{1, \dots, d\}$. Let N denote the number of concepts and let $\mathcal{C} = \{c_1, \dots, c_N\}$. Similarly, let $\mathcal{L}_1, \dots, \mathcal{L}_M$ denote the list of all proper *learning functions*, i.e., functions that map $X^m \times \{0,1\}^m$ to $\mathcal{C}$. A learning function $\mathcal{L}$ is said to be *consistent* if it maps every sample to a hypothesis from the corresponding version space.

If a proper deterministic learner ignores the parameters ε and δ in her input, she can be identified with a learning function (remember that we set aside computational issues). While the learners that we consider here in fact ignore ε and δ in their inputs, they are, however, randomized. Each-one of these can be identified with a probability distribution over the set of all learning functions. A randomized learner (or her strategy) is called *consistent* if she puts probability mass 1 on consistent learning functions.

3.3.1 Organization of This Section

The proof of existence of PAC-learners with universal constant performance ratios is quite involved. We will therefore present an overview of this section: Section 3.3.2 is devoted to learning under a fixed distribution P. This setting is cast as a zero-sum game between two players, the learner and her opponent, such that the Minimax Theorem from game theory applies. This leads to a nice characterization of $\delta_P^* = \delta_P^*(\varepsilon, m)$. It is furthermore shown that, when the opponent makes his draw first, there is a strategy for the learner that, although it does not at all depend on P, does not perform much worse than the best strategy that is specialized to P.

In the remaining sections, distribution P is not fixed, but chosen by an adversary, who picks a vector $\boldsymbol{z}$ of parameters from a set Δ and plays the distribution P_z. Ideally, we would like to design a learner who performs well in comparison to a learner with full prior knowledge of $P_{\boldsymbol{z}}$ for *all choices* of $\boldsymbol{z} \in \Delta$. This (as seen in Chapter 4) impossible endeavor is briefly described in terms of an appropriately defined zero-sum game in Section 3.3.3. After a short discussion of some measurability issues in Section 3.3.4, we pursue a less ambitious goal in Section 3.3.5, modeled again as a zero-sum game between the learner and her opponent, and show that there is a learner who performs well in comparison to a learner with full prior knowledge of $P_{\boldsymbol{z}}$ for *most choices* of $\boldsymbol{z} \in \Delta$. The total probability of the region containing the "bad choices" of $\boldsymbol{z}$ can be made as small as we like at the expense of a slightly degraded performance on the "good choices". Moreover, the underlying probability measure on Δ can be specified by a "user" (who can try to put high probability mass on realistic measures), and the strategy of the learner can be chosen such as to adapt to this specification.

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

1. The basic building stone is the matrix $\boldsymbol{A}_P^{\varepsilon,m}$, as defined in Section 3.3.2. The corresponding zero-sum game models PAC-learning under the fixed distribution P.

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

3. Since it is not possible to find a good strategy for the learner in the $\boldsymbol{R}$-game, we switch to the $\bar{\boldsymbol{R}}$-game. The payoff-matrix $\bar{\boldsymbol{R}}$ results from $\boldsymbol{R}$ basically by averaging over all $\boldsymbol{z} \in \Delta$. A learner playing successfully the $\bar{\boldsymbol{R}}$-game achieves a small "average performance ratio". We shall find a

good strategy for the learner in the $\bar{\boldsymbol{R}}$-game, and this will finally lead us to our main result.

3.3.2 Learning Under a Fixed Distribution Revisited

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

For every $\varepsilon > 0$, $i = 1, \dots, M$, $j = 1, \dots, N$, $\boldsymbol{x} \in X^m$, and $\boldsymbol{b} \in \{0,1\}^m$, let $\boldsymbol{I}_P^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$ be the Bernoulli variable indicating whether the hypothesis $\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b})$ is ε-inaccurate w.r.t. target concept c_j and domain distribution P, i.e.,

$$\boldsymbol{I}_P^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j] = \begin{cases} 1 & \text{if } P(\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b}) \neq c_j) > \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad . \tag{3.5}$$

Now, let

$$\boldsymbol{A}_P^{\varepsilon,m}[i,j] := \mathbb{E}_{\boldsymbol{x}\in P^m}\left[\boldsymbol{I}_P^{\varepsilon,\boldsymbol{x},c_j(\boldsymbol{x})}[i,j]\right] = \sum_{\boldsymbol{x}} P^m(\boldsymbol{x})\boldsymbol{I}_P^{\varepsilon,\boldsymbol{x},c_j(\boldsymbol{x})}[i,j] \tag{3.6}$$

$$= \Pr_{\boldsymbol{x}\sim P^m}\left(\mathcal{L}_i(\boldsymbol{x},c_j(\boldsymbol{x})) \text{ is } \varepsilon\text{-inaccurate for } c_j \text{ w.r.t. } P\right) \ . \tag{3.7}$$

If P, ε, m are obvious from context, we omit these letters as subscripts or superscripts in what follows. A randomized learner is given by a vector $\boldsymbol{p} \in [0,1]^M$ that assigns a probability p_i to every learning function $\mathcal{L}_i$ (so that $\sum_{i=1}^M p_i = 1$). Thus, we may identify learners with mixed strategies for the row-player in the zero-sum game associated with $\boldsymbol{A}$. We may view the column-player in this game as an opponent of the learner. A mixed strategy for the opponent is given by a vector $\boldsymbol{q} \in [0,1]^N$ that assigns an à-priori probability q_j to every possible target concept c_j (so that $\sum_{j=1}^N q_j = 1$). In the sequel, $\boldsymbol{A}_j$ denotes the j-th column of $\boldsymbol{A}$. If the learners plays strategy $\boldsymbol{p}$ and her opponent plays strategy $\boldsymbol{q}$ (or the pure strategy j, resp.), the former has to pay an amount of $\boldsymbol{p}^\top\boldsymbol{A}\boldsymbol{q}$ (or an amount of $\boldsymbol{p}^\top\boldsymbol{A}_j$, resp.) to the latter.

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

Recall the Minimax Theorem, which states that

$$\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top \boldsymbol{A}\boldsymbol{q} = \max_{\boldsymbol{q}} \min_{\boldsymbol{p}} \boldsymbol{p}^\top \boldsymbol{A}\boldsymbol{q} \ . \tag{3.8}$$

In the sequel, we denote the optimal value in (3.8) by $\delta_P^*(\varepsilon, m)$. In order to establish a relation between $\delta_P^*(\varepsilon, m)$ and strategies for learners without prior knowledge of P, we proceed as follows:

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

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

$$\tilde{\boldsymbol{A}}[i,j] = \begin{cases} 1 & \text{if } c_i \text{ is } \varepsilon\text{-inaccurate for } c_j \text{ w.r.t. } P \\ 0 & \text{otherwise} \end{cases} \ .$$

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

1. The opponent selects a vector $\boldsymbol{q} \in [0,1]^N$ specifying à-priori probabilities for the target concept. Note that this implicitly determines
   - the probability
     $$Q(\boldsymbol{b}|\boldsymbol{x}) = \sum_{j:c_j(\boldsymbol{x})=\boldsymbol{b}} q_j$$
     of labeling a given sample $\boldsymbol{x}$ by $\boldsymbol{b}$,
   - and the à-posteriori probabilities
     $$Q(j|\boldsymbol{x}, \boldsymbol{b}) = \begin{cases} \frac{q_j}{Q(\boldsymbol{b}|\boldsymbol{x})} & \text{if } c_j(\boldsymbol{x}) = \boldsymbol{b} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$
     for target concept c_j given the labeled sample $(\boldsymbol{x}, \boldsymbol{b})$.

   For sake of a compact notation, let $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ denote the vector whose j-th component is $Q(j|\boldsymbol{x}, \boldsymbol{b})$.

2. A labeled sample $(\boldsymbol{x}, \boldsymbol{b})$ is produced at random with probability
$$\Pr(\boldsymbol{x}, \boldsymbol{b}) = P^m(\boldsymbol{x})Q(\boldsymbol{b}|\boldsymbol{x}) \ . \tag{3.10}$$

3. The learner chooses a vector $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b}) \in [0,1]^N$ (that may depend on $P, \boldsymbol{q}$ and $(\boldsymbol{x}, \boldsymbol{b})$) specifying her mixed strategy w.r.t. payoff-matrix $\tilde{\boldsymbol{A}}$. We say that the learner's strategy $\tilde{\boldsymbol{p}}$ is *consistent* if, for every $(\boldsymbol{x}, \boldsymbol{b})$, $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})$ puts probability mass 1 on the version space for $(\boldsymbol{x}, \boldsymbol{b})$.

4. The learner suffers loss $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})^\top \tilde{\boldsymbol{A}} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ so that her expected loss, averaged over all labeled samples, evaluates to $\sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})^\top \tilde{\boldsymbol{A}} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$.

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

Lemma 11. *Let $\boldsymbol{q} \in [0,1]^N$ be an arbitrary but fixed mixed strategy for the learner's opponent. Then the following holds:*

1. *Every mixed strategy $\boldsymbol{p} \in [0,1]^M$ for the learner in the $\boldsymbol{A}$-game can be mapped to a mixed strategy $\tilde{\boldsymbol{p}}$ for the learner in the $\tilde{\boldsymbol{A}}$-game so that*
$$\boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} = \sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})^\top \tilde{\boldsymbol{A}} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \ . \tag{3.11}$$
Moreover, $\tilde{\boldsymbol{p}}$ is a consistent strategy for the $\tilde{\boldsymbol{A}}$-game if $\boldsymbol{p}$ is a consistent strategy for the $\boldsymbol{A}$-game.

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

Proof. Recall that M is the number of learning functions of the form $\mathcal{L} : X^m \times \{0,1\}^m \to \mathcal{C}$. Thus, every probability vector $\boldsymbol{p} \in [0,1]^M$ is a probability measure on the discrete product space $\Omega = \mathcal{C} \times \cdots \times \mathcal{C}$ with one factor $\mathcal{C}$ for every labeled sample $(\boldsymbol{x}, \boldsymbol{b}) \in X^m \times \{0,1\}^m$. Recall that $\mathcal{C} = \{c_1, \ldots, c_N\}$. Thus, every probability vector $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b}) \in [0,1]^N$ is a probability measure on the discrete space $\mathcal{C}$ which can be identified with factor $(\boldsymbol{x}, \boldsymbol{b})$ of Ω. We define the mapping $\boldsymbol{p} \mapsto \tilde{\boldsymbol{p}}$ by setting $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})$ equal to the marginal measure obtained by restricting $\boldsymbol{p}$ to factor $(\boldsymbol{x}, \boldsymbol{b})$ of Ω, i.e.,
$$\tilde{p}_{i'}(\boldsymbol{x}, \boldsymbol{b}) = \sum_{i:\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b})=c_{i'}} p_i \tag{3.12}$$

The following computation verifies (3.11):
$$\boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} \overset{(3.6)}{=} \sum_{\boldsymbol{x}} P^m(\boldsymbol{x}) \sum_{j=1}^{N} \sum_{i=1}^{M} \boldsymbol{I}^{\boldsymbol{x},c_j(\boldsymbol{x})}[i,j] p_i q_j$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{x},\boldsymbol{b}} P^m(\boldsymbol{x}) \sum_{j:c_j(\boldsymbol{x})=\boldsymbol{b}} \sum_{i=1}^{M} \boldsymbol{I}^{\boldsymbol{x},\boldsymbol{b}}[i,j] p_i q_j \\
&= \sum_{\boldsymbol{x},\boldsymbol{b}} P^m(\boldsymbol{x}) \sum_{j:c_j(\boldsymbol{x})=\boldsymbol{b}} \sum_{i'=1}^{N} \sum_{i:\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b})=c_{i'}} \underbrace{\boldsymbol{I}^{\boldsymbol{x},\boldsymbol{b}}[i,j]}_{=\tilde{\boldsymbol{A}}[i',j]} p_i q_j \\
&= \sum_{\boldsymbol{x},\boldsymbol{b}} P^m(\boldsymbol{x}) \sum_{j:c_j(\boldsymbol{x})=\boldsymbol{b}} \sum_{i'=1}^{N} \tilde{\boldsymbol{A}}[i',j] q_j \sum_{i:\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b})=c_{i'}} p_i \\
&\overset{(3.12)}{=} \sum_{\boldsymbol{x},\boldsymbol{b}} P^m(\boldsymbol{x}) \sum_{i'=1}^{N} \sum_{j:c_j(\boldsymbol{x})=\boldsymbol{b}} \tilde{\boldsymbol{A}}[i',j] \tilde{p}_{i'}(\boldsymbol{x},\boldsymbol{b}) q_j \\
&\overset{(3.9)}{=} \sum_{\boldsymbol{x},\boldsymbol{b}} P^m(\boldsymbol{x}) Q(\boldsymbol{b}|\boldsymbol{x}) \sum_{i'=1}^{N} \sum_{j=1}^{N} \tilde{\boldsymbol{A}}[i',j] \tilde{p}_{i'}(\boldsymbol{x},\boldsymbol{b}) Q(j|\boldsymbol{x},\boldsymbol{b}) \\
&\overset{(3.10)}{=} \sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x},\boldsymbol{b}) \tilde{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{b})^\top \tilde{\boldsymbol{A}} \tilde{\boldsymbol{q}}(\boldsymbol{x},\boldsymbol{b})
\end{aligned}
$$

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

$$p_i = \prod_{\boldsymbol{x},\boldsymbol{b}} \tilde{p}_{\mathcal{L}_i(\boldsymbol{x},\boldsymbol{b})}(\boldsymbol{x},\boldsymbol{b}) \ . \tag{3.13}$$

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

As far as consistency is concerned, finally note the following. If $\boldsymbol{p}$ puts probability mass 1 on consistent learning functions, then $\tilde{\boldsymbol{p}}$, defined according to (3.12), puts probability mass 1 on the version space for $(\boldsymbol{x},\boldsymbol{b})$. Conversely, if $\tilde{\boldsymbol{p}}$ puts probability mass 1 on the version space for $(\boldsymbol{x},\boldsymbol{b})$, then $\boldsymbol{p}$, defined according to (3.13), puts probability mass 1 on consistent learning functions. □

In the sequel, we list some consequences of Lemma 11. For instance, the lemma immediately implies that the optimal value in the $\boldsymbol{A}$-game (where $\boldsymbol{A} = \boldsymbol{A}_P^{\varepsilon,m}$) coincides with the optimal value in the $\tilde{\boldsymbol{A}}$-game (where $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}_P^{\varepsilon}$), i.e.,

$$
\begin{aligned}
\delta_P^*(\varepsilon,m) &= \min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top \boldsymbol{A} \boldsymbol{q} \\
&= \max_{\boldsymbol{q}} \left[\sum_{\boldsymbol{x},\boldsymbol{b}} \left(\Pr(\boldsymbol{x},\boldsymbol{b}) \cdot \min_{\tilde{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{y})} \left[\tilde{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{y})^\top \tilde{\boldsymbol{A}} \tilde{\boldsymbol{q}}(\boldsymbol{x},\boldsymbol{b}) \right] \right) \right] \ .
\end{aligned} \tag{3.14}
$$

Remark Notice that (3.14) offers the opportunity to prove (non-constructively) the *existence* of a good learning strategy for the $\boldsymbol{A}$-game with the learner making the first draw, by presenting a good (sample-dependent) learning strategy for the $\tilde{\boldsymbol{A}}$-game with the learner making the second draw.

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

Corollary 12. *If $\delta_P^*(\varepsilon, m) = 0$, then the following holds: for every $\boldsymbol{x} \in X^m$ such that $P^m(\boldsymbol{x}) > 0$, and for every $\boldsymbol{b} \in \{0,1\}^m$, there exists a hypothesis $h^* \in \mathcal{C}$ that is ε-accurate for every hypothesis in the version space for $(\boldsymbol{x}, \boldsymbol{b})$.*

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

$$\forall h^* \in \mathcal{C}, \exists c \in \mathcal{C} : \ c(\boldsymbol{x}) = \boldsymbol{b} \ \wedge \ P(h^* \neq c) > \varepsilon \ . \tag{3.15}$$

We have to show that this assumption leads to the conclusion $\delta_P^*(\varepsilon, m) > 0$. To this end, pick $\boldsymbol{x}$ and $\boldsymbol{b}$ such that $P^m(\boldsymbol{x}) > 0$ and such that (3.15) is valid. Let $\boldsymbol{q}$ be the uniform distribution on $\mathcal{C}$. According to (3.10), we may conclude that $\Pr(\boldsymbol{x}, \boldsymbol{b}) > 0$. An inspection of (3.14) now reveals that $\delta_P^*(\varepsilon, m) > 0$, as desired. □

Corollary 13. *If $\delta_P^*(\varepsilon, m) = 0$, then every consistent learner suffers loss 0 in the $\boldsymbol{A}_P^{2\varepsilon,m}$-game.*

Proof. According to Lemma 11, it suffices to show that every strategy $\tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})$ for the $\tilde{\boldsymbol{A}}$-game that puts probability mass 1 on the version space for $(\boldsymbol{x}, \boldsymbol{b})$ suffers loss 0 in the $\tilde{\boldsymbol{A}}$-game. According to Corollary 12, there exists a hypothesis h^* that is ε-accurate for every hypothesis in the version space $V := V_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{b})$. This clearly implies, that every hypothesis in V is 2ε-accurate for every other function in V. Thus, putting probability mass 1 on V leads to loss 0. □

We close this section with a result that prepares the ground for our analysis of general PAC-learners in Section 3.3:

Lemma 14. *Let $\varepsilon > 0$ be a given accuracy, and let $m \geq 1$ be a given sample size. For every probability vector $\boldsymbol{q} \in [0,1]^N$, and every domain distribution P, the following holds:*

$$\sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})^\top \tilde{\boldsymbol{A}}_P^{2\varepsilon} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \leq 2\delta_P^*(\varepsilon, m) \tag{3.16}$$

Proof. The left hand-side in (3.16) represents the learners loss in the $\tilde{\boldsymbol{A}}_P^{2\varepsilon}$-game when the opponent plays strategy $\boldsymbol{q}$ (the à-priori probabilities for the target concepts) and the learner plays strategy $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ on sample $(\boldsymbol{x}, \boldsymbol{b})$. Recall that

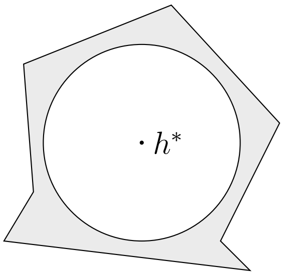


Figure 3.1: The polygon represents the version space for $(\boldsymbol{x}, \boldsymbol{b})$. The circle represents the hypotheses that are ε-close to h^* with respect to P. The shaded area represents the loss induced by h^* on sample $(\boldsymbol{x}, \boldsymbol{b})$.

$\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ is the collection of à-posteriori probabilities for the target concepts. Since the à-posteriori probabilities outside the version space $V := V_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{b})$ are zero, only target concepts in V can contribute to the learner's loss. In the remainder of the proof, we simply write $\tilde{\boldsymbol{A}}^{\varepsilon}$ instead of $\tilde{\boldsymbol{A}}^{\varepsilon}_P$, and $\tilde{\boldsymbol{A}}^{\varepsilon}_i$ denotes the i-th row of this matrix. Given $\boldsymbol{q}$ and $(\boldsymbol{x}, \boldsymbol{b})$, the term $\tilde{\boldsymbol{A}}^{\varepsilon}_i \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ represents the loss suffered by a learner with hypothesis c_i in the $\tilde{\boldsymbol{A}}^{\varepsilon}_P$-game. This loss equals the total à-posteriori-probability of the hypotheses from the version space that are *not* ε-close to c_i w.r.t. domain distribution P. It is minimized by picking a hypothesis $h^* = c_{i^*(\boldsymbol{x},\boldsymbol{b})} \in \mathcal{C}$ which maximizes the total à-posteriori probability of hypotheses that *are* ε-close to h^* w.r.t. P. With this notation, it follows that

$$\sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{A}}^{\varepsilon}_{i^*(\boldsymbol{x},\boldsymbol{b})} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \le \delta^*_P(\varepsilon, m) \tag{3.17}$$

with equality if the strategy $\boldsymbol{q}$ of the learner's opponent is optimal. The situation is visualized in Figure 3.1. We are now prepared to verify (3.16). Assume that, given $\boldsymbol{q}$ and $(\boldsymbol{x}, \boldsymbol{b})$, the learner applies strategy $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ instead of choosing the best hypothesis h^*. There are two "unlucky cases":

- The learners random hypothesis falls into the shaded area in Figure 3.1.
- The opponent's random target concept falls into this shaded area.

Each of the unlucky events happens with a probability that equals the total à-posteriori probability of the hypotheses in the shaded area, i.e. it happens with probability $\tilde{\boldsymbol{A}}^{\varepsilon}_{i^*(\boldsymbol{x},\boldsymbol{b})} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$, respectively. If none of the unlucky events occurs, then the learner's hypothesis *and* the target concept fall into the circle in Figure 3.1 so that they are 2ε-close to each other w.r.t. P. Our discussion shows that

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

which, in view of (3.17), yields (3.16). □

It is important to note that no knowledge of P is required to play the strategy $\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$ in the $\tilde{\boldsymbol{A}}$-game (with the opponent making the first draw). Nevertheless, as made precise in Lemma 14, this is a reasonably good strategy for *any* fixed underlying domain distribution. Note furthermore that a learner with strategy $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$ is consistent, since the à-posteriori probabilities assign probability mass 1 to the version space for $(\boldsymbol{x}, \boldsymbol{b})$.

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

- a probability measure μ on the d-dimensional probability simplex (recall that $X = \{1, \ldots, d\}$)

  $$\Delta := \{\boldsymbol{z} \in [0,1]^d : \ z_1 + \cdots + z_d = 1\} \ , \tag{3.18}$$

  where every $\boldsymbol{z} \in \Delta$ represents a measure $P_{\boldsymbol{z}}$ such that $P_{\boldsymbol{z}}(x) = z_x$, i.e., putting probability mass z_x on the x-th element from X, for $x = 1, \ldots, d$,

- *conditional* à-priori probabilities $\boldsymbol{q}(j|\boldsymbol{z})$ for choosing target concept c_j provided that $P_{\boldsymbol{z}}$ is the underlying domain distribution.

In this general setting, the à-posteriori probabilities, associated with the hypotheses after having seen a labeled sample $(\boldsymbol{x}, \boldsymbol{b})$, previously denoted $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b})$, are now conditioned on $\boldsymbol{z}$ as well and therefore denoted $\tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z})$. Consequently, the loss suffered by a learner playing strategy $\tilde{\boldsymbol{p}}$ formally looks as follows (compare with (3.14)):

$$\mathbb{E}_{\boldsymbol{z}\sim\mu}\left[\sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z}) \tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b})^\top \tilde{\boldsymbol{A}}^{2\varepsilon}_{P_{\boldsymbol{z}}} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z})\right] \tag{3.19}$$

According to Lemma 14, the learner can favorably play strategy $\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$ in the fixed-distribution setting. As can be seen from (3.19), there is no uniform good choice for $\tilde{\boldsymbol{p}}$ anymore in the general setting since $\tilde{\boldsymbol{q}}$ is conditioned to $\boldsymbol{z}$. For this reason, we are not done yet.

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

3.3.3 Towards a Small Worst Performance Ratio

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

Recall from the previous section that

$$\boldsymbol{A}_{P_{\boldsymbol{z}}}^{2\varepsilon,m}[i,j] = \Pr_{\boldsymbol{x}\sim P_{\boldsymbol{z}}^m}\left(\mathcal{L}_i(\boldsymbol{x}, c_j(\boldsymbol{x})) \text{ is } 2\varepsilon\text{-inaccurate for } c_j \text{ w.r.t. } P_{\boldsymbol{z}}\right) \ .$$

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

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

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

$$\frac{\max_{j\in[N]} \boldsymbol{p}^\top \boldsymbol{A}_j^{(\boldsymbol{z})}}{\delta_{\boldsymbol{z}}^*}$$

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

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

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

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

is called the *worst performance ratio* of strategy $\boldsymbol{p}$.[4] The value

$$\rho^*(\varepsilon,m) := \min_{\boldsymbol{p}} \rho^{\boldsymbol{p}}(\varepsilon,m)$$

is the best possible worst performance ratio. Note that it coincides with the smallest loss suffered by the learner in the game with payoff-matrix $\boldsymbol{R}$. A very strong result would be $\rho^*(\varepsilon,m) \le k$ for some constant k. Since this is impossible, as we will see in Chapter 4, we shall pursue a weaker goal in the following.

[4] Note that, according to Corollary 13, $\delta_{\boldsymbol{z}}^* = 0$ implies that, for every consistent strategy $\boldsymbol{p}$, $\boldsymbol{p}^\top \boldsymbol{A}_j^{(\boldsymbol{z})}/\delta_{\boldsymbol{z}}^* = 0/0 = 1$.

3.3.4 Some Measurability Issues

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

- f is measurable if and only if
$$\forall \alpha \in \mathbb{R} : \{\omega \in \Omega : \ f(\omega) \geq \alpha\} \in \mathcal{A} \ . \tag{3.20}$$
- The sum and the product of two measurable functions is measurable.
- The reciprocal of a strictly positive (or strictly negative, resp.) measurable function is measurable.
- The infimum (or supremum, resp.) of a sequence of measurable functions is measurable.
- If Ω is a Borel set equipped with the Borel-algebra and $f(\omega)$ is a continuous function in $\omega \in \Omega$, then f is measurable.

These closure properties have the following implications:

Corollary 15. *Let Ω be a Borel set equipped with the Borel-algebra. Let K be a finite or countably infinite set and, for every $k \in K$, let Ω_k be a Borel-set such that $\Omega = \cup_{k \in K} \Omega_k$. Let $f(\omega)$ be a numerical function in $\omega \in \Omega$ that is continuous on every Ω_k. Then f is measurable.*

Proof. For every $\alpha \in \mathbb{R}$, consider the decomposition
$$\{\omega \in \Omega : \ f(\omega) \geq \alpha\} = \bigcup_{k \in K} \{\omega \in \Omega_k : \ f(\omega) \geq \alpha\} \ .$$

Since f is continuous on every Ω_k, the right hand-side is a countable union of Borel-sets and therefore a Borel-set itself. Thus, f satisfies (3.20) and is a measurable function. □

Corollary 16. *Let the d-dimensional probability simplex Δ be equipped with the Borel-algebra. Then, for every choice of $\varepsilon, m, \boldsymbol{x}, \boldsymbol{b}, i, j$, $\boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$, $\boldsymbol{A}_{P_{\boldsymbol{z}}}^{\varepsilon,m}[i,j]$, and $\delta_{\boldsymbol{z}}^*$ are measurable functions in $\boldsymbol{z} \in \Delta$.*

Proof. Since
$$\boldsymbol{A}_{P_{\boldsymbol{z}}}^{\varepsilon,m}[i,j] = \sum_{\boldsymbol{x}} P_{\boldsymbol{z}}^m(\boldsymbol{x}) \boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},c_j(\boldsymbol{x})}[i,j] \text{ and } \delta_{\boldsymbol{z}}^* = \min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top \boldsymbol{A}_{P_{\boldsymbol{z}}}^{\varepsilon,m} \boldsymbol{q} \ ,$$

it suffices to show that $\boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$ is a measurable function in $\boldsymbol{z} \in \Delta$. For any $X' \subseteq X$, with $X = \{1, \ldots, d\}$, consider the hyperplane

$$H_{X'} := \left\{ \boldsymbol{z} : \sum_{x \in X'} z_x = \varepsilon \right\}$$

and the halfspaces

$$H_{X'}^{+} := \left\{ \boldsymbol{z} \in \Delta : \sum_{x \in X'} z_x > \varepsilon \right\} , \; H_{X'}^{-} := \left\{ \boldsymbol{z} \in \Delta : \sum_{x \in X'} z_x \leq \varepsilon \right\} .$$

These halfspaces decompose Δ into finitely many cells where every single cell can be written as an intersection of Δ with finitely many halfspaces (which clearly yields a Borel-set). An inspection of (3.5) shows that the discontinuities of $\boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$ occur on the hyperplane

$$H_{X'} \text{ such that } X' = \{x \in X : \; \mathcal{L}_i(\boldsymbol{x}, \boldsymbol{b})(x) \neq c_j(x)\}$$

only and that $\boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$ is continuous on every single cell of the mentioned decomposition. According to Corollary 15, $\boldsymbol{I}_{P_{\boldsymbol{z}}}^{\varepsilon,\boldsymbol{x},\boldsymbol{b}}[i,j]$ is a measurable function in $\boldsymbol{z} \in \Delta$. □

3.3.5 The Average Performance Ratio

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

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

Note that these sets are Borel-sets since $\delta_{\boldsymbol{z}}^*$ is a measurable function according to Corollary 16. Since probability measures are continuous from above, we get

$$\lim_{\zeta \to 0} \mu(\Delta_1(\zeta)) = \mu\left(\bigcap_{\zeta > 0} \Delta_1(\zeta) \right) = \mu(\emptyset) = 0 \; . \qquad (3.21)$$

In our discussion of PAC-learning, it is justified to focus on domain distributions $P_{\boldsymbol{z}}$ such that $\boldsymbol{z} \in \Delta_2(\zeta)$ for the following reasons:

- Distributions $P_{\boldsymbol{z}}$ such that $\boldsymbol{z} \in \Delta_0$ do not pose any problem to a consistent learner. Compare with Corollary 13.

- Distributions $P_{\boldsymbol{z}}$ such that $\boldsymbol{z} \in \Delta_1(\zeta)$ might pose a problem to PAC-learners but the probability mass assigned to them by μ can be made arbitrarily small according to (3.21).

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

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

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

$$\boldsymbol{R}^{(\boldsymbol{z})}[i,j] = \frac{1}{\delta^*_{P_{\boldsymbol{z}}}(\varepsilon,m)} \cdot \boldsymbol{A}^{2\varepsilon,m}_{P_{\boldsymbol{z}}}[i,j]$$

is bounded by $1/\zeta$ and measurable according to Corollary 16. The quantity

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

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

$$\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\top \bar{\boldsymbol{R}}^{\zeta} \boldsymbol{q} = \max_{\boldsymbol{q}} \min_{\boldsymbol{p}} \boldsymbol{p}^\top \bar{\boldsymbol{R}}^{\zeta} \boldsymbol{q} \tag{3.22}$$

Clearly, the optimal value in (3.22) coincides with the best possible average performance ratio. Equation (3.22) offers the opportunity to (non-constructively) show the *existence* of a “good” learning strategy with the learner making the first draw, by presenting a “good” learning strategy with the learner making the second draw, where “good” here means “achieving a small average performance ratio”. (Compare with the remark to Equation (3.14).)

Lemma 17. *For every mixed strategy $\boldsymbol{q}$ of the opponent in the $\bar{\boldsymbol{R}}^{\zeta}$-game, there exists a consistent mixed strategy $\boldsymbol{p}$ for the learner such that $\boldsymbol{p}^\top \bar{\boldsymbol{R}}^{\zeta} \boldsymbol{q} \leq 2$.*

Proof. From the decomposition (3.11), we get the following decomposition of $\boldsymbol{p}^\top \bar{\boldsymbol{R}}^{\zeta} \boldsymbol{q}$:

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

Here, $\delta_{\boldsymbol{z}}^* = \delta_{P_{\boldsymbol{z}}}^*(\varepsilon, m)$, $\Pr(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z}) = P_{\boldsymbol{z}}^m(\boldsymbol{x})Q(\boldsymbol{b}|\boldsymbol{x})$, $\tilde{\boldsymbol{A}}^{(\boldsymbol{z})} = \tilde{\boldsymbol{A}}_{P_{\boldsymbol{z}}}^{2\varepsilon}$, and the quantities $Q(\boldsymbol{b}|\boldsymbol{x}), \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}$ are derived from $\boldsymbol{q}$ and $\boldsymbol{p}$, respectively, as explained in Section 3.3.2. According to Lemma 14 (applied to $P = P_{\boldsymbol{z}}$),

$$\sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z}) \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{A}}^{(\boldsymbol{z})} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \le 2\delta_{\boldsymbol{z}}^* \ . \tag{3.23}$$

According to Lemma 11, there exists a mixed strategy $\boldsymbol{p}$ for the learner such that $\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$. With this choice of $\boldsymbol{p}$, we get

$$\boldsymbol{p}^\top \bar{R}^\zeta \boldsymbol{q} = \mathbb{E}_{\boldsymbol{z} \sim \mu'} \left[\frac{1}{\delta_{\boldsymbol{z}}^*} \sum_{\boldsymbol{x},\boldsymbol{b}} \Pr(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{z}) \tilde{\boldsymbol{p}}(\boldsymbol{x}, \boldsymbol{b}) \tilde{\boldsymbol{A}}^{(\boldsymbol{z})} \tilde{\boldsymbol{q}}(\boldsymbol{x}, \boldsymbol{b}) \right] \overset{(3.23)}{\le} 2 \ ,$$

as desired. □

Corollary 18. *For every probability measure μ on Δ, there exists a consistent mixed strategy for the learner with an average performance ratio of at most 2.*

Proof. The Minimax Theorem (applied to the submatrix of $\bar{\boldsymbol{R}}^\zeta$ with rows corresponding to consistent learning functions) combined with Lemma 17 yields the result. □

Corollary 18 talks about the average performance ratio which, by definition, is the performance ratio averaged over $\Delta_2(\zeta)$ only. Furthermore, it does not explicitly bound the probability mass of domain distributions $P_{\boldsymbol{z}}$ for which the supervised learner performs considerably worse than the learner with full prior knowledge of $P_{\boldsymbol{z}}$. The next result, the main result in this chapter, fills these gaps:

Theorem 19. *For every probability measure μ on Δ and for every $k > 1$, $\gamma > 0$, there exists a mixed strategy $\boldsymbol{p}$ for the learner such that, for $j = 1, \ldots, N$,*

$$\mu\left(\left\{\boldsymbol{z} \in \Delta : \ \boldsymbol{p}^\top \boldsymbol{R}_j^{(\boldsymbol{z})} \ge 2k\right\}\right) < \frac{1}{k} + \gamma \ .$$

Proof. For sake of brevity, let $E := \{\boldsymbol{z} \in \Delta : \ \boldsymbol{p}^\top \boldsymbol{R}_j^{(\boldsymbol{z})} \ge 2k\}$. With this notation, $\mu(E)$ is bounded from above by

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

The first term contributes 0 because of Corollary 13. The second-one contributes at most $\mu(\Delta_1(\zeta))$ which is smaller than γ for every sufficiently small ζ. The third-one contributes at most $1/k$ according to Corollary 18 combined with Markov's inequality. Thus, $\Pr_{\boldsymbol{z} \sim \mu}(E) < 1/k + \gamma$, as desired. □

According to Theorem 19, there exists a mixed strategy $\boldsymbol{p}$ for a learner without any prior knowledge of the domain distribution such that, in comparison to the best learner with full prior knowledge of the domain distribution, a performance ratio of $2k$ is achieved for the “vast majority” of distributions. The total probability mass of distributions (measured according to μ) not belonging to the “vast majority” is bounded by $1/k + \gamma$ (where γ may be chosen as small as we like). So Theorem 19 is the result that we had announced in the introduction.

3.4 Conclusions and Open Questions

The most pressing open questions will be answered in the next chapter. However, there are still several interesting problems:

- The performance ratio for classes of finite VC-dimension in Theorem 9 has a (logarithmic) dependence on ε. We do not know whether it possible to remove that dependence.
- Theorem 9 is only valid for classes that contain the constant zero- and the constant one-function. It is an open question whether this requirement can be dropped.
- For every class over a finite domain, we have shown the mere existence of a learner whose average performance ratio is bounded by 2. It would be interesting to gain more insight how this strategy actually works and under which conditions it can be implemented efficiently.
- In the second part of the present chapter, we restricted ourselves to finite domains X for ease of technical exposition (e.g., compactness of the $|X|$-dimensional probability simplex). We conjecture however that this restriction can be weakened considerably.
- In the present chapter, we restricted ourselves to the realizable setting. It is an open question whether corresponding results hold in more relaxed settings, like agnostic learning or learning with noise.

Chapter 4

Where Unlabeled Data Does Help

This chapter is based on the paper "Unlabeled Data Does Provably Help" [DSS13][1], which is a joint work of the author with Hans Ulrich Simon and Balázs Szörényi. All results in this chapter were originally published in this paper.

4.1 Introduction

The existing analysis of the semi-supervised setting can be summarized roughly as follows:

- The benefit of unlabeled samples can be enormous if the target concept and the domain distribution satisfy some suitable extra assumptions (see [BB10]).
- On the other hand, the benefit seems to be marginal if we do not impose any extra-assumptions (see Chapter 3 and [BDLP08]).

These findings perfectly match with the common belief expressed in conjecture (2.2), that an extra assumption, i.e., some kind of compatibility between the target concept and the domain distribution, is needed for adding horsepower to semi-supervised algorithms. However, the results of the second type are not yet fully convincing:

- The paper [BDLP08] provides some upper bounds on the label complexity in the fully supervised setting and some lower bounds, that match up to a small constant factor, in the semi-supervised setting (or even in the setting with a distribution P that is known to the learner). These bounds however are established only for some special concept classes over the real line. It is unclear whether they generalize to a broader variety of concept classes.

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

- Theorem 9 in Chapter 3 shows that the conjecture is true for arbitrary finite classes and (up to logarithmic factors) for classes of finite VC-dimension. This of course leaves open the case of classes with an infinite VC-dimension. Furthermore, it would be interesting to know if the performance ratio in Theorem 9 must depend on $\mathcal{C}$.

- In the second part of Chapter 3 we analyzed arbitrary concept classes over finite domains and, in Theorem 19, showed the existence of a purely supervised learning algorithm with a universal constant performance ratio for the "vast majority" of domain distributions. This however does not exclude the possibility that there still exist "bad distributions" leading to a poor performance of the learner.

In this chapter, we reconsider the question whether unlabeled samples can be of significant help to a learner even when we do not impose any extra-assumptions on the PAC-learning model. A comparably old paper, [DKRZ94], indicates that an affirmative answer to this question is thinkable (despite of the fact that it was written a long time before semi-supervised learning became an issue). In [DKRZ94] it is shown that there exists a concept class $\mathcal{C}_\infty$ and a family $\mathcal{P}_\infty$ of domain distributions such that the following holds:

1. For each $P \in \mathcal{P}_\infty$, $\mathcal{C}_\infty$ is properly PAC-learnable under the fixed distribution P.

2. $\mathcal{C}_\infty$ is *not properly* PAC-learnable under unknown distributions taken from $\mathcal{P}_\infty$.

These results point into the right direction for our purpose, but they are not precisely what we want:

- Although "getting a large unlabeled sample" comes close to "knowing the domain distribution", it is not quite the same. (In fact, one can show that $\mathcal{C}_\infty$, with domain distributions taken from $\mathcal{P}_\infty$, is *not* PAC-learnable in the semi-supervised setting.)

- The authors of [DKRZ94] do *not* show that $\mathcal{C}_\infty$ is *not PAC-learnable* under unknown distributions P taken from $\mathcal{P}_\infty$. In fact, their proof uses a target concept that almost surely (w.r.t. P) assigns 1 to every instance in the domain. But the (proper!) learner must not return the constant one-function of error 0 because of her commitment to hypotheses from $\mathcal{C}_\infty$.

4.1.1 Main Results

Can we generalize Theorem 9 to concept classes $\mathcal{C}$ of infinite VC-dimension provided that the domain distribution is taken from a family $\mathcal{P}$ such that

$m^*_{\mathcal{C},P}(\varepsilon,\delta) < \infty$ for all $P \in \mathcal{P}$? This question will be answered in the negative by the following result (proved in Section 4.3):

Theorem 20. *There exists a concept class $\mathcal{C}_*$ over domain $\{0,1\}^*$ and a family $\mathcal{P}_*$ of domain distributions such that the following holds:*

1. *There exists a semi-supervised algorithm L such that, for all $P \in \mathcal{P}_*$,*

$$m^L_{\mathcal{C}_*,P}(\varepsilon,\delta) = O\left(\frac{1}{\varepsilon^2} + \frac{\ln(1/\delta)}{\varepsilon}\right) \ .$$

   This implies the same upper bound on $m^*_{\mathcal{C}_*,P}$ for all $P \in \mathcal{P}_*$.

2. *For every fully supervised algorithm L and for all $\varepsilon < 1/2, \delta < 1$:*

$$\sup_{P\in\mathcal{P}_*} m^L_{\mathcal{C}_*,P}(\varepsilon,\delta) = \infty \ .$$

Does there exist a universal constant performance ratio k (not depending on $\mathcal{C}$) such that we get a result similar to Theorem 9 but with $k(\mathcal{C})$ replaced by k? Is it possible that the “bad domain distributions” occurring in Theorem 19 are avoidable and just an artifact of our analysis? The following result (proved in Section 4.3) shows that, even for classes of finite VC-dimension, such a universal constant does not exist and that the bad distributions in Theorem 19 can not be avoided:

Theorem 21. *There exists a sequence $(\mathcal{C}_n)_{n\geq 1}$ of concept classes over domains $(\{0,1\}^n)_{n\geq 1}$ such that $\lim_{n\to\infty} \mathrm{VCdim}(\mathcal{C}_n) = \infty$ and a sequence $(\mathcal{P}_n)_{n\geq 1}$ of domain distribution families such that the following holds:*

1. *There exists a semi-supervised algorithm L that PAC-learns $(\mathcal{C}_n)_{n\geq 1}$ under any unknown distribution and, for all $P \in \mathcal{P}_n$,*

$$m^L_{\mathcal{C}_n,P}(\varepsilon,\delta) = O\left(\frac{1}{\varepsilon^2} + \frac{\ln(1/\delta)}{\varepsilon}\right) \ .$$

   This implies the same upper bound on $m^*_{\mathcal{C}_n,P}$ for all $P \in \mathcal{P}_n$.

2. *For every fully supervised algorithm L and all $\varepsilon < 1/2, \delta < 1$:*

$$\sup_{n\geq 1, P\in\mathcal{P}_n} m^L_{\mathcal{C}_n,P}(\varepsilon,\delta) = \infty \ .$$

Some comments are in place here:

- Since the class $\mathcal{C}_*$ from Theorem 20 has a countable domain, namely $\{0,1\}^*$, $\mathcal{C}_*$ occurs (via projection) as a subclass in every concept class that shatters a set of infinite cardinality. A similar remark applies to the sequence $(\mathcal{C}_n)_{n\geq 1}$ and concept classes that shatter finite sets of arbitrary size. Thus every concept class of infinite VC-dimension contains subclasses that are significantly easier to learn in the semi-supervised setting of the PAC-model (in comparison to the full supervised setting).

- An error bound $\varepsilon = 1/2$ is trivially achieved by random guesses for the unknown label. Let α and β be two arbitrary small, but strictly positive, constants. Theorems 20 and 21 imply that even the modest task of returning, with a success probability of at least α, a hypothesis of error at most $1/2 - \beta$ cannot be achieved in the fully supervised setting unless the number of labeled examples becomes arbitrarily large.

- Theorem 20 proves that the conjecture (2.2) by Ben-David et al. is false. Indeed, the theorem shows that even the performance ratio between the sample complexity of fully supervised and semi-supervised learners (instead of learners in the fixed distribution setting, as considered by Ben-David et al.), is not always finite. Furthermore, Theorem 21 implies that the results from [BDLP08] for simple classes on the real line do *not* generalize to arbitrary finite classes, unless we allow the performance ratio to become arbitrarily large.

- $\mathcal{C}_n$ is not an artificially constructed or exotic class: it is in fact the class of non-negated literals over n boolean variables, which occurs as a subset of many popular concept classes (e.g., monomials, decision lists, halfspaces). The class $\mathcal{C}_*$ is a natural generalization of $\mathcal{C}_n$ to the set of boolean strings of arbitrary length.

- The classes $\mathcal{C}_*, \mathcal{P}_*$ from Theorem 20, defined in Section 4.3, are close relatives of the classes $\mathcal{C}_\infty, \mathcal{P}_\infty$ from [DKRZ94], but the adversary argument that we have to employ is much more involved than the corresponding argument in [DKRZ94] (where the learner was assumed to be proper and had been fooled mainly because of her commitment to hypotheses from $\mathcal{C}_\infty$).

4.2 Prerequisites From Probability Theory

Let X be an integer-valued random variable. As usual, a most likely value a for X is called a *mode* of X. In this chapter, the largest integer that is a mode of X is denoted $\mathrm{mode}(X)$. As usual, X is said to be *unimodal* if $\Pr(X = x)$ is increasing with x for all $x \leq \mathrm{mode}(X)$, and decreasing with x for all $x \geq \mathrm{mode}(X)$.

Let Ω be a space equipped with a σ-algebra of events and with a probability measure P. For any sequence $(A_n)_{n\geq 1}$ of events, $\limsup_{n\to\infty} A_n$ is defined as the set of all $\omega \in \Omega$ that occur in infinitely many of the sets A_n, i.e., $\limsup_{n\to\infty} A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m$. We briefly remind the reader of the Borel-Cantelli Lemma:

Lemma 22 ([Fel68]). *Let $(A_n)_{n\geq 1}$ be a sequence of independent events, and let $A = \limsup_{n\to\infty} A_n$. Then $P(A) = 1$ if $\sum_{n=1}^{\infty} P(A_n) = \infty$, and $P(A) = 0$ otherwise.*

Corollary 23. *Let $(A_n)_{n\geq 1}$ be a sequence of independent events with the property $\sum_{n=1}^{\infty} P(A_n) = \infty$ and let $B_{k,n}$ be the set of all $\omega \in \Omega$ that occur in at least k of the events $A_1, \ldots, A_n$. Then, for any $k \in \mathbb{N}$, $\lim_{n\to\infty} P(B_{k,n}) = 1$.*

Proof. Note that $B_{k,n} \subseteq B_{k,n+1}$ for every n. Since probability measures are continuous from below, it follows that $\lim_{n\to\infty} P(B_{k,n}) = P(\cup_{n=1}^{\infty} B_{k,n})$. Since, obviously, $\limsup_{n\to\infty} A_n \subseteq \cup_{n=1}^{\infty} B_{k,n}$, an application of the Borel-Cantelli Lemma yields the result. □

The following result, which is a variant of the Central Limit Theorem for triangular arrays, is known in the literature as the Lindeberg-Feller Theorem:

Theorem 24 ([Chu74]). *Let $(X_{n,i})_{n\in\mathbb{N}, i\in[n]}$ be a (triangular) array of random variables such that*

1. $\mathbb{E}[X_{n,i}] = 0$ *for all* $n \in \mathbb{N}$, $i = 1, \ldots, n$.
2. $X_{n,1}, \ldots, X_{n,n}$ *are independent for every* $n \in \mathbb{N}$.
3. $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[X_{n,i}^2] = \sigma^2 > 0$.
4. *For each* $\varepsilon > 0$, $\lim_{n\to\infty} s_n(\varepsilon) = 0$ *where* $s_n(\varepsilon) = \sum_{i=1}^{n} \mathbb{E}[X_{n,i}^2 \mathbb{I}(|X_{n,i}| \geq \varepsilon)]$.

Then it holds that

$$\lim_{n\to\infty} P\Big(a < \frac{1}{\sigma} \cdot \sum_{i=1}^{n} X_{n,i} < b\Big) = \varphi(b) - \varphi(a) \ ,$$

where φ denotes the density function of the standard normal distribution.

An easy padding argument shows that this theorem holds "mutatis mutandis" for triangular arrays of the form $(X_{n_k,i})$ where $i = 1, \ldots, n_k$ and $(n_k)_{k\geq 1}$ is an increasing and unbounded sequence of positive integers. (The limes is then taken for $k \to \infty$.) We furthermore note that, for the special case of independent Bernoulli variables $X_{n,i}$ with probability p_i of success, Theorem 24 applies to the triangular array $(X_{n,i} - p_i)/\sigma_n$ where $\sigma_n^2 = \sum_{i=1}^{n} p_i(1-p_i)$ with $\lim_{n\to\infty} \sigma_n^2 = \infty$. A similar remark applies to the more general case of bounded random variables.

The following result is an immediate consequence of Theorem 24 (plus the remarks thereafter):

Lemma 25. *Let $l(k) = o(\sqrt{k})$. Let $(n_k)_{k\geq 1}$ be an increasing and unbounded sequence of positive integers. Let $(p_{k,i})_{k\in\mathbb{N},i\in[n_k]}$ range over all triangular arrays of parameters in $[0,1]$ such that*

$$\forall k \in \mathbb{N} : \sum_{i=1}^{n_k} p_{k,i}(1-p_{k,i}) \geq k \ . \tag{4.1}$$

Let $(X_{k,i})_{k\in\mathbb{N},i\in[n_k]}$ be the corresponding triangular array of row-wise independent Bernoulli variables. Then the function h given by

$$h(k) = \sup_{(p_{k,i})} \sup_{s\in\{0,\ldots,n_k\}} P\left(\Big|\sum_{i=1}^{n_k} X_{k,i} - s\Big| < l(k)\right)$$

approaches 0 as k approaches infinity.

Proof. Assume for sake of contradiction that $\limsup_{k\to\infty} h(k) > 0$. Then there exist $(p_{k,i})$ satisfying (4.1) and $s_k \in \{0,\ldots,n_k\}$ such that

$$\limsup_{k\to\infty} P\left(\Big|\sum_{i=1}^{n_k} X_{k,i} - s_k\Big| < l(k)\right) > 0 \ . \tag{4.2}$$

The random variable $S_k = \sum_{i=1}^{n_k} X_{k,i}$ has mean $\mu_k = \sum_{i=1}^{n_k} p_{k,i}$ and variance $\sigma_k^2 = \sum_{i=1}^{n_k} p_{k,i} \cdot (1-p_{k,i}) \geq k$. The Lindeberg-Feller Theorem applied to the triangular array $\left(\frac{X_{k,i}-p_i}{\sigma_k}\right)$ yields

$$\lim_{k\to\infty} P\left(a < \frac{S_k - \mu_k}{\sigma_k} < b\right) = \varphi(b) - \varphi(a) \ . \tag{4.3}$$

For S_k to hit a given interval of length $2l(k)$ (like the interval $[s_k - l(k), s_k + l(k)]$ in (4.2)) it is necessary for $(S_k - \mu_k)/\sigma_k$ to hit a given interval of length $2l(k)/\sigma_k$. Note that $\lim_{k\to\infty} l(k)/\sigma_k = 0$ because $\sigma_k \geq \sqrt{k}$ and $l(k) = o(\sqrt{k})$. Thus the hitting probability approaches 0 as k approaches infinity. This contradicts to (4.2). □

For ease of later reference, we let $k(\beta)$ for $\beta > 0$ be a function such that $h(k) \leq \beta$ for all $k \geq k(\beta)$. Note that such a function must exist according to Lemma 25.

Corollary 26. *With the notation and assumptions from Lemma 25, the following holds: the probability mass of the mode of $\sum_{i=1}^{n_k} X_{k,i}$ is at most β for all $k \geq k(\beta)$.*

The following result implies the unimodality of binomially distributed random variables:

Lemma 27 ([KG71]). *Every sum of independent Bernoulli variables (with possibly different probabilities of success) is unimodal.*

4.3 Proof of Theorems 20 and 21

Throughout this section, we set $X_n = \{0,1\}^n$ and $X_* = \{0,1\}^*$. We will identify a finite string $x \in X_*$ with the infinite string that starts with x and ends with an infinite sequence of zeros. $\mathcal{C}_*$ denotes the family of functions $c_i : X_* \to \{0,1\}$, $i \in \mathbb{N} \cup \{0\}$, given by $c_0(x) = 0$ and $c_i(x) = x_i$ for all $i \geq 1$. Note that $c_i(x) = 0$ for all $i > |x|$. $\mathcal{C}_n$ denotes the class of functions obtained by restricting a function from $\mathcal{C}_*$ to the subdomain X_n. For every $i \geq 1$, let $p_i = 1/\log_2(3+i)$. For every permutation σ of $1,\ldots,n$, let P_σ be the probability measure on X_n obtained by setting $x_{\sigma(i)} = 1$ with probability p_i (resp. $x_{\sigma(i)} = 0$ with probability $1-p_i$) independently for $i = 1,\ldots,n$. $\mathcal{P}_n = \{P_\sigma\}$ denotes the family of all such probability measures on X_n. Note that P_σ can also be considered as a probability measure on X_* (that is centered on X_n). $\mathcal{P}_*$, a family of probability measures on X_*, is defined as $\cup_{n\geq 1}\mathcal{P}_n$.

Lemma 28. 1. *$\mathcal{C}_*$ is properly PAC-learnable under any fixed distribution $P_\sigma \in \mathcal{P}_*$ from $O(1/\varepsilon^2 + \ln(1/\delta)/\varepsilon)$ labeled examples.*

2. *For any (unknown) $P_\sigma \in \mathcal{P}_*$, $\mathcal{C}_*$ is properly PAC-learnable in the semi-supervised setting from $O(\ln(n/\delta)/\varepsilon)$ unlabeled and $O(1/\varepsilon^2 + \ln(1/\delta)/\varepsilon)$ labeled examples. Here, n denotes the smallest index such that $P_\sigma \in \mathcal{P}_n$.*

3. *There exists a semi-supervised algorithm L that PAC-learns $\mathcal{C}_n$ under any unknown domain distribution. Moreover, for all $P \in \mathcal{P}_n$, $m^L_{\mathcal{C}_n,P}(\varepsilon,\delta) = O(1/\varepsilon^2 + \ln(1/\delta)/\varepsilon)$.*

Proof. 1. Let σ be a permutation of $1,\ldots,n$. For all $i > n$ it holds $c_i = \emptyset$ almost surely w.r.t. P_σ. For all $2^{2/\varepsilon} - 3 \leq i \leq n$ it holds $P_\sigma(c_{\sigma(i)} \neq \emptyset) = P_\sigma(c_{\sigma(i)}) = p_i \leq \varepsilon/2$. Thus, setting $N = \lceil 2^{2/\varepsilon}\rceil - 4$, $\{\emptyset, c_{\sigma(1)}, \ldots, c_{\sigma(N)}\}$ forms an $\varepsilon/2$-covering of $\mathcal{C}_*$ with respect to P_σ. An application of Theorem 6 now yields the result.

2. The very first unlabeled example reveals the parameter n such that the unknown measure P_σ is centered on X_n. Note that, for every $i \in [n]$, $x_i = 1$ with probability $p_{\sigma^{-1}(i)}$. It is an easy application of the Chernoff-bound (combined with the Union-bound) to see that $O(\ln(n/\delta)/\varepsilon)$ unlabeled examples suffice to retrieve (with probability $1-\delta/2$ of success) an index set $I \subset [n]$ with the following properties: on one hand, I includes all $i \in [n]$ such that $p_{\sigma^{-1}(i)} \geq \varepsilon/2$. On the other hand, I excludes all $i \in [n]$ such that $p_{\sigma^{-1}(i)} \leq \varepsilon/8$. Consequently $\{\emptyset\} \cup \{c_i : i \in I\}$ is an $\varepsilon/2$-covering of $\mathcal{C}_n$ with respect to P_σ and its size is bounded by $1 + |I| \leq 2^{8/\varepsilon}$. Another application of Theorem 6 now yields the result.

3. The third part in Lemma 28 is an immediate consequence of Theorem 7 and the fact that, as proved above, $N_{\mathcal{C}_n}(\varepsilon/6) = 2^{O(1/\varepsilon)}$ (regardless of the value of n). □

Lemma 29. *Let L be a fully supervised algorithm designed to PAC-learn $\mathcal{C}_*$ under any unknown distribution taken from $\mathcal{P}_*$. For every finite sample size m and for all $\alpha, \beta > 0$, an adversary can achieve the following: with a probability of at least $1-\alpha$ the hypothesis returned by L has an error of at least $1/2-\beta$.*[2]

Proof. The proof will run through the following stages:

1. We first fix some technical notations and conditions (holding in probability) which the proof builds on.
2. Then we specify the strategy of the learner's adversary.
3. We argue that, given the strategy of the adversary, the learner has probably almost no advantage over random guesses.
4. We finally verify the technical conditions.

Let us start with Stage 1. (Though somewhat technical it will help us to provide a precise description of the subsequent stages.) Let $\boldsymbol{M} \in \{0,1\}^{(m+1)\times(\mathbb{N}\setminus\{1\})}$ be a random matrix (with columns indexed by integers not smaller than 2) such that the entries are independent Bernoulli variables where the variable $\boldsymbol{M}[i,j]$ has probability $p_j = 1/\log_2(3+j) < 1/2$ of success. Let $\boldsymbol{M}(n)$ denote the finite matrix composed of the first $n-1$ columns of $\boldsymbol{M}$. Let $k = \max\{\lceil 1/\alpha\rceil, k(2\beta)\}$ where $k(\beta)$ is the function from the remark right after Lemma 25. In Stage 4 of the proof, we will show that there exists $n = n_k \in \mathbb{N}$ such that, with probability at least $1-1/k$, the following conditions are valid for each bit pattern $\boldsymbol{b} \in \{0,1\}^{m+1}$:

(A) $\boldsymbol{b} \in \{0,1\}^{m+1}$ coincides with at least $4k^2$ columns of $\boldsymbol{M}(n)$.

(B) Let $\boldsymbol{b}' \in \{0,1\}^m$ be the bit pattern obtained from $\boldsymbol{b}$ by omission of the final bit. Call column $j \geq 2$ of $\boldsymbol{M}(n)$ "marked" if its first m bits yield pattern $\boldsymbol{b}'$. Let $I \subseteq \{2,\ldots,n\}$ denote the set of indices for marked columns. Then, $\sum_{i\in I} p_i \geq 2k$ so that $\sum_{i\in I} p_i(1-p_i) \geq k$ (because $p_i < 1/2$).

The strategy of the adversary (Stage 2 of the proof) is as follows: he sets $n = n_k$, picks a permutation σ of $1,\ldots,n$ uniformly at random, chooses domain distribution P_σ, and selects the target concept $c^* = c_t$ such that $t = \sigma(1)$. In the sequel, probabilities are simply denoted $P(\cdot)$. Note that the component x_t of an example x can be viewed as a fair coin since $P(x_t = 1) = p_1 = 1/\log_2(4) = 1/2$. The learning task resulting from this setting is related to the technical definitions and conditions from Stage 1 as follows:

[2] Loosely speaking, the learner has "probably almost no advantage over random guesses".

- The first m rows of the matrix $\boldsymbol{M}(n)$ are the components $\sigma(2), \ldots, \sigma(n)$ of the m labeled examples.
- The bits of $\boldsymbol{b}' \in \{0,1\}^m$ are the t-th components of the m labeled examples. These bits are perfectly random, and they are identical to the classification labels.
- The set $I \subseteq \{2, \ldots, n\}$ points to all marked columns of $\boldsymbol{M}(n)$, i.e., it points to all columns of $\boldsymbol{M}(n)$ which are duplicates of $\boldsymbol{b}'$.
- Row $m{+}1$ of $\boldsymbol{M}$ represents an unlabeled test point that has to be classified by the learner.

The adversary passes also the set $J = \{\sigma(i) : i \in I \cup \{1\}\}$, with the understanding that index t of the target concept is an element of J, and the set $I \subseteq \{2, \ldots, n\}$ as additional information to the learner. This maneuver marks the end of Stage 2 in our proof.

We now move on to Stage 3 of the proof and explain why the strategy of the adversary leads to a poor learning performance (thereby assuming that conditions (A) and (B) hold). Note that, by symmetry, every index in J has the same à-posteriori probability to coincide with t. Because the learner has no way to break the symmetry between the indices in J before she sees the test point x, the best prediction for the label of x does not depend on the individual bits in x but only on the number of ones in the bit positions from J, i.e., it only depends on the value of

$$Y' = \sum_{j \in J} x_j = x_{\sigma(1)} + \sum_{i \in I} x_{\sigma(i)} = x_{\sigma(1)} + Y$$

where

$$Y = \sum_{i \in I} x_{\sigma(i)} = \sum_{i \in I} \boldsymbol{M}[m+1, i] \enspace .$$

Note that the learner knows the distribution of Y (given by the parameters $(p_i)_{i \in I}$) since the set I had been passed on to her by the adversary. For sake of brevity, let $\ell = x_{\sigma(1)}$ denote the classification label of the test point x. Given a value s of Y' (and the fact that the à-priori probabilities for $\ell = 0$ and $\ell = 1$ are equal), the Bayes-decision is in favor of the label $\ell \in \{0,1\}$ which maximizes $P(Y' = s|\ell)$. Clearly, $P(Y' = s|\ell = 1) = P(Y = s-1)$ and $P(Y' = s|\ell = 0) = P(Y = s)$. Thus, the Bayes-decision is in favor of $\ell = 0$ if and only if $P(Y = s) \geq P(Y = s-1)$. Since Y is a sum of independent Bernoulli variables, we may apply Lemma 27 and conclude that Y has a unimodal distribution. It follows that the Bayes-decision is the following threshold function: be in favor of $\ell = 0$ iff $Y' \leq \mathrm{mode}(Y)$. The punchline of this discussion is as follows: the Bayes-decision is independent of the true label ℓ

unless Y hits its mode (so that $Y' = Y + \ell$ is either $\mathrm{mode}(Y)$ or $\mathrm{mode}(Y) + 1$). It follows that the Bayes-error is at least $(1 - P(Y = \mathrm{mode}(Y)))/2$. Because of Condition (B) and the fact that $k \geq k(2\beta)$, we may apply Corollary 26 and obtain $P(Y = \mathrm{mode}(Y)) \leq 2\beta$ so that the Bayes-error is at least $1/2 - \beta$.
We finally enter Stage 4 of the proof and show that conditions (A) and (B) hold with a probability of at least $1 - \alpha$ provided that $n = n_k$ is large enough. Let $\boldsymbol{b}$ range over all bit patterns from $\{0,1\}^{m+1}$. Consider the events

- $A_r(\boldsymbol{b})$: $\boldsymbol{b} \in \{0,1\}^{m+1}$ coincides with the r-th column of $\boldsymbol{M}$.
- $B_{k,n}(\boldsymbol{b})$: $\boldsymbol{b} \in \{0,1\}^{m+1}$ coincides with at least k columns of $\boldsymbol{M}(n)$.

It is easy to see that $\sum_{r=1}^{\infty} P(A_r(\boldsymbol{b})) = \infty$. Applying the Borel-Cantelli Lemma to the events $(A_r(\boldsymbol{b}))_{r \geq 1}$ and Corollary 23 to the events $(B_{4k^2,n}(\boldsymbol{b}))_{k,n \geq 1}$, we arrive at the following conclusion. There exists $n_k(\boldsymbol{b}) \in \mathbb{N}$ such that, for all $n \geq n_k(\boldsymbol{b})$, the probability of $B_{4k^2,n}(\boldsymbol{b})$ is at least $1 - 1/(2^{m+3}k)$. We set $n = n_k = \max_{\boldsymbol{b}} n_k(\boldsymbol{b})$. Then, the probability of $B_{4k^2,n} = \bigcap_{\boldsymbol{b} \in \{0,1\}^{m+1}} B_{4k^2,n}(\boldsymbol{b})$ is at least $1 - 1/(4k)$. In other words: with a probability of at least $1 - 1/(4k)$, each $\boldsymbol{b} \in \{0,1\}^{m+1}$ coincides with at least $4k^2$ columns of $\boldsymbol{M}(n)$. Thus condition (A) is violated with a probability of at most $1/(4k)$.
We move on to condition (B). With $p = \sum_{i \in I} p_i$, we can decompose $P(B_{4k^2,n})$ as follows:

$$P(B_{4k^2,n}) = P\left(B_{4k^2,n} \mid p < 2k\right) \cdot P\left(p < 2k\right) + P\left(B_{4k^2,n} \mid p \geq 2k\right) \cdot P\left(p \geq 2k\right)$$

Note that, according to the definitions of $B_{4k^2,n}(\boldsymbol{b})$ and $B_{4k^2,n}$, event $B_{4k^2,n}$ implies that $Y \geq 4k^2$ because there must be at least $4k^2$ occurrences of 1 in row $m+1$ and in the marked columns of $\boldsymbol{M}(n)$. On the other hand, $\mathbb{E}[Y] = p$. According to Markov's inequality, $P\left(Y \geq 4k^2 | p < 2k\right) \leq (2k)/(4k^2) = 1/(2k)$. Thus, $P(B_{4k^2,n}) \leq 1/(2k) + P\left(p \geq 2k\right)$. Recall that, according to condition (A), $1 - 1/(4k) \leq P(B_{4k^2,n})$. Thus, $P\left(p \geq 2k\right) \geq 1 - 1/(4k) - 1/(2k) = 1 - 3/(4k)$. Since $k \geq 1/\alpha$, we conclude that the probability to violate one of the conditions (A) and (B) is bounded by $1/k \leq \alpha$. □

We are now ready to complete the proofs of our main results. Theorem 20 is a direct consequence of the second statement in Lemma 28 and of Lemma 29. The first part of Theorem 21 is a direct consequence of the third statement in Lemma 28. As for the second part, an inspection of the proof of Lemma 29 reveals that the adversary argument uses a "finite part" $\mathcal{C}_n$ of $\mathcal{C}_*$ only (with n chosen sufficiently large).

4.4 Conclusions and Open Questions

As we have seen in this chapter, it is impossible to show in full generality that unlabeled samples have a marginal effect only in the absence of any extra assumptions. The following open questions remain:

- It would be interesting to explore which distributions are similar in this respect to the artificial distribution families $\mathcal{P}_*$ and $\mathcal{P}_n$ that were discussed in this chapter.
- We would like to know if the bounds of Theorem 9 are tight (either for special classes or for the general case).
- Like in the previous chapter, we restricted ourselves to the realizable setting. Again, it would be interesting to extend our results to more relaxed settings, like the agnostic learning or learning with noise.

Chapter 5

A Case Study: Co-training and the Conditional Independence Assumption

This chapter is based on the following two articles:

- “Supervised Learning and Co-training” [DSS14][1], which is a joint work of the author with Hans Ulrich Simon and Balázs Szörényi
- “The Optimal PAC Bound for Intersection-closed Concept Classes” [Dar15][1], published solely by the author of this thesis

The results in this chapter were originally published in the two articles given above, with the following exceptions: the results from Section 5.3.2 and Section 5.5 were previously unpublished; the main results from the part “Combinatorial bounds for classes with small (co-)singleton sizes” in Section 5.4.3 and from Section 5.4.5, which is a joint work with Balázs Szörényi, were mentioned in the former article, but remained unproven (i.e, the relevant proof details, such as definitions, lemmas and—of course—the proofs themselves, were not published before).

5.1 Introduction

As mentioned in Chapter 2, 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, expressed in an extra assumption. As the results in Chapter 3 have shown, this intuition is supported in so far as for most classes that are relevant in practice 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).

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

In this chapter, 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 [BM98].

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.

Note that a similar study for the popular Cluster Assumption was done by Singh, Nowak and Zhu in [SNZ08]. They show that the value of unlabeled data under their formalized Cluster Assumption varies with the minimal margin between clusters.

5.1.1 Co-training and the Conditional Independence Assumption

The results from Chapter 3 give a justification of using extra assumptions in the semi-supervised framework in order to make real use of having access to unlabeled data. Co-training under the Conditional Independence Assumption aims at this direction. The co-training model was introduced by Blum and Mitchell in [BM98], 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 5.2.) A theoretical analysis of Co-training under the Conditional Independence Assumption [BM98], and the weaker α-expanding Assumption [BBY04], was accomplished by Balcan and Blum in [BB10]. They work in Valiant's model of PAC-learning [Val84] 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]

This chapter complements their results: we also work in the PAC model and prove sample complexity bounds, but in our case the learner has no access to

[2] This is one of the results which impressively demonstrate the striking potential of properly designed semi-supervised learning strategies although the underlying extra assumptions are somewhat idealized and therefore not likely to be strictly satisfied in practice. See [BBY04, WZ10] for suggestions of relaxed assumptions.

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 sample 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 5.1.3 for more details.

5.1.2 Related Work

This chapter touches several related fields besides of Semi-supervised learning and Co-training. Let us review these connections:

Agnostic active learning We make extensive use of a suitably defined variant of Hanneke's disagreement coefficient. (See Section 5.3.4 for a comparison of the two notions.) The disagreement coefficient was first used in learning theory, without calling it by this name, by Giné and Koltchinskii in a paper on agnostic learning in 2006 [GK06], which was based on the work of Alexander [Ale86, Ale87]. It was independently discovered by Hanneke in 2007 [Han07] to analyze the sample complexity of agnostic active learning.

To our knowledge the work presented in this chapter is, besides a remark about classical PAC-learning in Hanneke's thesis [Han09] (which we will improve upon in Section 5.3.5), 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 [GG89]. 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.

PAC-learning intersection-closed classes On a side-note, we will answer an open question of Auer and Ortner [AO07] in Section 5.3.5 in the positive.

Since Valiant introduced the PAC-learning framework in [Val84], it is an open problem to give sharps bounds on the number of labeled examples necessary to successfully learn in the realizable setting. Obviously, the best known worst-case lower and upper bounds (see Theorems 4 and 5) leave a gap of $\ln(1/\varepsilon)$.

Warmuth conjectured in [War04] that the factor of $\ln(1/\varepsilon)$ in the upper bound can be overcome. He furthermore conjectured that this can be achieved by the *one-inclusion graph algorithm*, which—in the case of concept classes that are closed under intersection—collapses to the *closure algorithm*, whose output is the smallest consistent hypothesis.

For some special intersection-closed classes, which possess an additional combinatorial property, Auer and Ortner provided a proof of the conjecture in [AO07]. We will show, by using the disagreement coefficient, that Warmuth's conjecture is indeed true for *all* intersection-closed classes.

To our knowledge this is the first proof of sharp bounds on the sample complexity for a natural family of classes in the realizable PAC-learning framework. As a related result, Balcan and Long [BL13] could recently provide sharp bounds for efficiently learning halfspaces under special domain distributions.

5.1.3 Main Results

This chapter is a continuation of the line of research from the first chapters, aiming at investigating the problem: how much can the learner benefit from knowing the underlying distribution. We 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 (while we also provide experimental results, the experiments are conducted on artificial data only).

As mentioned above, the model of Co-training under the Conditional Independence Assumption was introduced in [BM98] as a setting where semi supervised can be superior to fully supervised learning. Indeed, in [BB10] it was shown that a single labeled example suffices for PAC-learning in the realizable case if unlabeled data is available. Recall Theorem 4, which states that supervised, realizable PAC-learning without any extra assumption requires d/ε labeled examples (up to logarithmic factors) where d denotes the VC-dimension of the concept class and ε is the accuracy parameter. The step from d/ε to just a single labeled example is a giant one. In this chapter, 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 $\Omega(d/\varepsilon)$ from Theorem 5 on the number of examples needed by a traditional (not co-trained) PAC-learner.

5.1.4 Organization of This Chapter

The remainder of the chapter is structured as follows.

Section 5.2 clarifies the notations and formal definitions that are used throughout the chapter and mentions some elementary facts.

In Section 5.3 we study a suitably defined variant of Hanneke's disagreement coefficient. The results of Section 5.3 seem to have implications for active learning and might be of independent interest. We now provide a more detailed view of the content of Section 5.3:

- Section 5.3.1 presents a fundamental inequality that relates our variant of Hanneke's disagreement coefficient [Han07] to a purely combinatorial parameter, $s(\mathcal{C})$, which is closely related to the "unique negative dimension" from [GG89]. 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 $s(\mathcal{C})$.
- Furthermore, in Sections 5.3.1 and 5.3.2 we look into the question how to choose a hypothesis class $\mathcal{H}$ in such a way that the afore-mentioned product of the VC-dimension and our disagreement coefficient is minimized.
- Section 5.3.3 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.
- Section 5.3.4 compares our variant of the disagreement coefficient with Hanneke's definition.
- Section 5.3.5 contains a remark in which we shortly leave the framework of Co-training and answer an open question about PAC-learning intersection closed classes.

In Section 5.4 we can finally give the our main results concerning the sample complexity of supervised Co-training. Again, a more detailed look at this section is in order:

- Section 5.4.1 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 5.3).

- Section 5.4.2 shows that all general upper bounds from Section 5.4.1 are (nearly) tight.
- Section 5.4.3 gives (almost) matching upper and lower bounds in completely combinatorial terms. Interestingly, the learning strategy that is best from the perspective of a worst-case analysis has one-sided error.
- Section 5.4.4 presents improved bounds for classes with special properties.
- Section 5.4.5 considers a generalization of the framework to sample points given by k-tuples.
- Section 5.4.6 shows a negative result in the more relaxed model of Co-training with α-expansion.

Section 5.5 provides experimental results on artificial data generated according to the worst-case distribution used to prove the lower bound on the sample complexity.

The closing Section 5.6 contains some final remarks and open questions.

5.2 Preliminaries – Co-training and the Disagreement Coefficient

In Co-training [BM98], it is assumed that there is a pair of concept classes, $\mathcal{C}_1$ and $\mathcal{C}_2$, and that random sample points come in pairs $(x_1, x_2) \in X_1 \times X_2$. Moreover, the domain distribution P, according to which the random sample points are generated, is perfectly compatible with the target concepts, say $c_1^* \in \mathcal{C}_1$ and $c_2^* \in \mathcal{C}_2$, in the sense that $c_1^*(x_1) = c_2^*(x_2)$ with probability 1. (For this reason, we sometimes denote the target label as $c^*(x_1, x_2)$.) As in [BM98, BB10], 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 [BB10], we perform a PAC-style analysis of Co-training under the Conditional Independence Assumption. But unlike [BB10], 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".

Definition 30. *Let $\mathcal{C}$ be a concept class over domain X and $\mathcal{H} \supseteq \mathcal{C}$ a hypothesis class over the same domain.*

For any $V \subseteq \mathcal{C}$, the disagreement region *of V is given by*

$$DIS(V) := \{x \in X : \exists h, h' \in V \text{ s.t. } h(x) \neq h'(x)\} \ .$$

We define the following variants of disagreement coefficients:

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

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

$$\theta(\mathcal{C},\mathcal{H}) := \sup_{P,c^*} \theta(\mathcal{C},\mathcal{H}|P,c^*)$$

In the degenerate case that both denominator and numerator are zero, we define $\theta(\mathcal{C}|P,X',c^*) := 1$. *For sake of brevity, we omit* $\mathcal{H}$ *in the case* $\mathcal{H} = \mathcal{C}$*, e.g.,* $\theta(\mathcal{C}) := \theta(\mathcal{C},\mathcal{C})$.

Note that

$$\theta(\mathcal{C},\mathcal{H}) \leq \theta(\mathcal{C}) \leq |\mathcal{C}| - 1 \ . \tag{5.1}$$

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

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

and an application of the union bound. Moreover, in the case $|\mathcal{C}| \geq 2$ it holds that

$$1 \leq \theta(\mathcal{C},\mathcal{H}) \ , \tag{5.2}$$

which can be seen by considering distributions P whose support is a subset of the disagreement region (note that a non-empty disagreement region always exists for $X' = \emptyset$, if $|\mathcal{C}| \geq 2$).

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

Example 31. *Let* X *be* $\mathbb{R}^2$ *and let* $\mathcal{C} = \mathcal{H}$ *be the class of closed homogeneous half planes. The target concept* c^* *is illustrated in Figure 5.1 by the hatched 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* c^**, is attached to each sample point, which we represent in Figure 5.1 by "*$+$*" and "*$-$*".*

$V_{\mathcal{C}}(X',c^*)$*, 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* $V_{\mathcal{C}}(X',c^*)$*, 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 shaded area, on which* h *and* c^* *differ, measured by* P*.*

The disagreement region, which consists of all points where some *hypothesis from* $V_{\mathcal{C}}(X',c^*)$ *disagrees with the target* c^**, is shown in the right-hand part of Figure 5.1.*

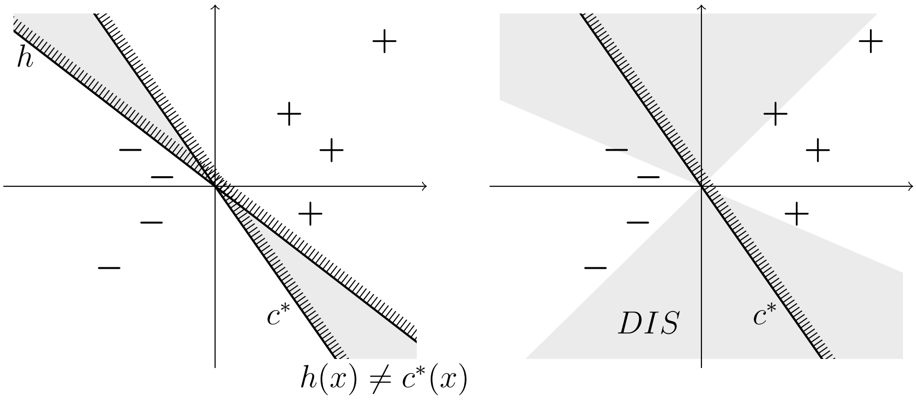


Figure 5.1: An illustration of c^*, X', h, the error of h (left-hand side) and the disagreement region (right-hand side) for Example 31. The hatched sides indicate the half planes c^* and h.

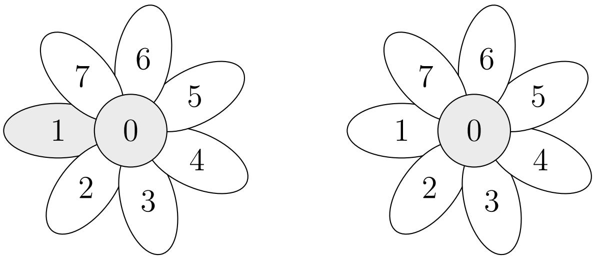


Figure 5.2: This drawing depicts the concepts $\{0, 1\}$ and $\{0\}$ in SF_7. Each concept in SF_n consists of the kernel 0 and at most one of the petals from 1 to n. The class is named after the sunflower and fulfills the well-known definition of "sunflower" in combinatorics, but is otherwise unrelated.

The disagreement coefficient $\theta(\mathcal{C}, \mathcal{H}|P, X', c^*)$ is the ratio of the size of the disagreement region, measured again by P, and the largest possible error of a half plane consistent with the sample.

We will now calculate θ for the following class, which will be useful for proving lower bounds in section 5.4.2:

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

See Figure 5.2 for a visualization of SF_n.

Lemma 32. $\theta(\mathrm{SF}_n) = n$.

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

Thus,

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

Conversely, $\theta(SF_n) \leq |\mathrm{SF}_n| - 1 = n$ (according to (5.1)). □

The main usage of this disagreement coefficient is as follows. First note that we have $P(DIS(V_{\mathcal{C}}(X', c^*))) \leq \theta(\mathcal{C}, \mathcal{H}) \cdot \sup_{h \in V_{\mathcal{H}}(X', c^*)} P(h \neq c^*)$ for every choice of P, X', c^*. This inequality holds in particular when X' consists of m points in X chosen independently at random according to P. According to the classical sample size bound of Theorem 4, there exists a sample size $m = \tilde{O}(\mathrm{VCdim}(\mathcal{H})/\varepsilon)$ such that, with probability at least $1 - \delta$, $\sup_{h \in V_{\mathcal{H}}(X', c^*)} P(h \neq c^*) \leq \varepsilon$. Thus, with probability at least $1 - \delta$ (taken over the set of random sample points X'), $P(DIS(V_{\mathcal{C}}(X', c^*))) \leq \theta(\mathcal{C}, \mathcal{H}) \cdot \varepsilon$. This discussion is summarized in the following lemma:

Lemma 33. *There exists a sample size $m = \tilde{O}(\mathrm{VCdim}(\mathcal{H})/\varepsilon)$ such that the following holds for every probability measure P on domain X and for every target concept $c^* \in \mathcal{C}$. With probability $1 - \delta$, taken over a random sample set X' of size m, $P(DIS(V_{\mathcal{C}}(X', c^*))) \leq \theta(\mathcal{C}, \mathcal{H}) \cdot \varepsilon$.*

This lemma indicates that one should choose $\mathcal{H}$ so as to minimize $\theta(\mathcal{C}, \mathcal{H}) \cdot \mathrm{VCdim}(\mathcal{H})$. Note that making $\mathcal{H}$ more powerful leads to smaller values of $\theta(\mathcal{C}, \mathcal{H})$ but comes at the price of an increased VC-dimension. See Sections 5.3.1 and 5.3.2 for a discussion about the optimal choice of $\mathcal{H}$.

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 $c^* \in \mathcal{C}$, there exists $h \in V_{\mathcal{H}}(X', c^*)$ such that $h(x) = 0$ ($h(x) = 1$, resp.) for every $x \in DIS(V_{\mathcal{C}}(X', c^*))$. 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 40 and its proof.

Lemma 34. *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 P, X', c^*. Let h_{min} be the hypothesis in $V_{\mathcal{H}}(X', c^*)$ that errs on positive examples of c^* only, and let h_{max} be the hypothesis in $V_{\mathcal{H}}(X', c^*)$ that errs on negative examples of c^* only. We conclude that $DIS(V_{\mathcal{C}}(X', c^*)) \subseteq \{x : h_{min}(x) \neq h_{max}(x)\}$. From this and the triangle inequality, it follows that

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

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

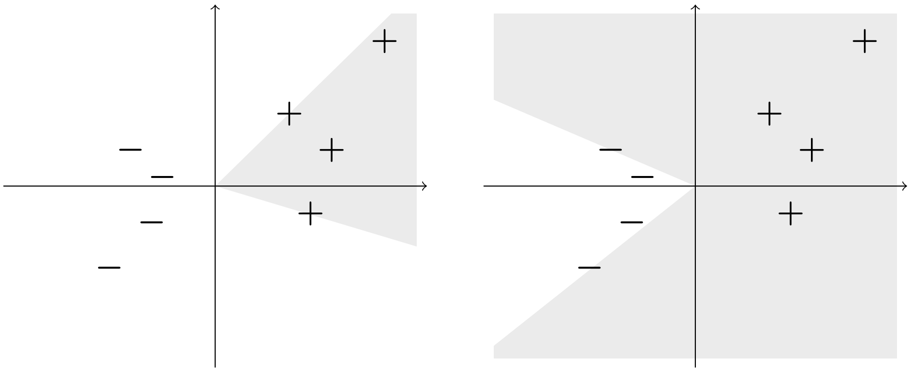


Figure 5.3: The shaded areas represent hypotheses with plus- and minus-sided errors. Note that the disagreement region, as seen in Figure 5.1, lies inside of the minus-sided hypothesis and outside of the plus-sided one.

Example 35. *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 $\theta(\mathrm{POWERSET}) \leq 2$ and $\theta(\mathrm{HALFINTERVALS}) \leq 2$.*

Let the class $\mathcal{C}$ consist of both the open and the closed homogeneous half planes, let $\mathcal{H}$ be the class of unions and intersections of two half planes from $\mathcal{C}$ and let $X = \mathbb{R}^2 \setminus \{\mathbf{0}\}$. It is easy to see that $\mathcal{H}$ contains hypotheses with plus-sided errors (the smallest pie slice with apex at $\mathbf{0}$ that includes all positive examples in a sample; see Figure 5.3) and hypotheses with minus-sided errors (the complement of the smallest pie slice with apex at $\mathbf{0}$ that includes all negative examples in a sample) w.r.t. $\mathcal{C}$. Thus, $\theta(\mathcal{C}, \mathcal{H}) \leq 2$. Note that $\mathcal{H}$ is neither closed under intersection nor closed under union.

5.3 A Closer Look at the Disagreement Coefficient

5.3.1 A Combinatorial Upper Bound

Definition 36. *Let $s^+(\mathcal{C})$ denote the largest number of instances in X such that every binary pattern on these instances with exactly one "+"-label can be realized by a concept from $\mathcal{C}$. In other words: $s^+(\mathcal{C})$ denotes the cardinality of the largest singleton subclass[3] of $\mathcal{C}$. If $\mathcal{C}$ contains singleton subclasses of arbitrary size, we define $s^+(\mathcal{C})$ as infinite.*

[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 $h' \in \mathcal{C}'$. The singleton class of size n is defined as the set $\mathcal{C}_{\mathrm{sng}} = \{\{x\} : x \in X\} = \{\{1\}, \{2\}, \dots, \{n\}\}$ over the domain $X = \{1, \dots, n\}$.

		X			
		1	2	3	4
	c_1	$+$	$-$	$-$	$-$
$\mathcal{C}_{\text{sng}}$	c_2	$-$	$+$	$-$	$-$
	c_3	$-$	$-$	$+$	$-$
	c_4	$-$	$-$	$-$	$+$

		X			
		1	2	3	4
	c_1	$-$	$+$	$+$	$+$
$\mathcal{C}_{\text{co}-\text{sng}}$	c_2	$+$	$-$	$+$	$+$
	c_3	$+$	$+$	$-$	$+$
	c_4	$+$	$+$	$+$	$-$

Figure 5.4: The class of singletons $\mathcal{C}_{\text{sng}}$ and co-singletons $\mathcal{C}_{\text{co}-\text{sng}}$ over the domain $X = \{1, 2, 3, 4\}$.

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.

See the left-hand side of Figure 5.4 for an illustration of a singleton class.

Lemma 37. *$\mathcal{C} \subseteq \mathcal{C}^+$, $\mathcal{C}^+$ 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 34, $\theta(\mathcal{C}, \mathcal{C}^+) \leq 2$. □

We aim at a similar result that holds for arbitrary (not necessarily intersection-closed) concept classes. To this end, we proceed as follows.

Definition 38. *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,*

[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 $h' \in \mathcal{C}'$. The co-singleton class of size n is defined as the set $\mathcal{C}_{\text{co}-\text{sng}} = \{X \setminus \{x\} : x \in X\}$ over the domain $X = \{1, \ldots, n\}$.

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.

Furthermore, we define

$$s(\mathcal{C}) := 4 \cdot \max\{s^+(\mathcal{C}), s^-(\mathcal{C})\} \ .$$

See the right-hand side of Figure 5.4 for an illustration of a co-singleton class.

By duality, Lemma 37 translates into the following result:

Corollary 39. *$\mathcal{C} \subseteq \mathcal{C}^-$, $\mathcal{C}^-$ 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 40. *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* $\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C}, \mathcal{H}) \leq s(\mathcal{C})$*.*

Proof. $\mathcal{C} \subseteq \mathcal{H}$ is obvious. The bound on the VC-dimension is obtained as follows. If m instances are given, then, by Lemma 37 and Corollary 39, 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 [Sau72]. 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 34 and the fact that, because of Lemma 37 and Corollary 39, $\mathcal{H} = \mathcal{C}^+ \cup \mathcal{C}^-$ 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 [GG89] as the "unique negative dimension" of $\mathcal{C}$. He showed that it characterizes PAC-learnability from positive examples alone.

Example 41. *Let us continue with the classes from Example 35. 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 $s^+(\mathit{POWERSET}) = s^-(\mathit{POWERSET}) = n$, while the class of half intervals only satisfies $s^+(\mathit{HALFINTERVALS}) = s^-(\mathit{HALFINTERVALS}) = 1$. 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.*

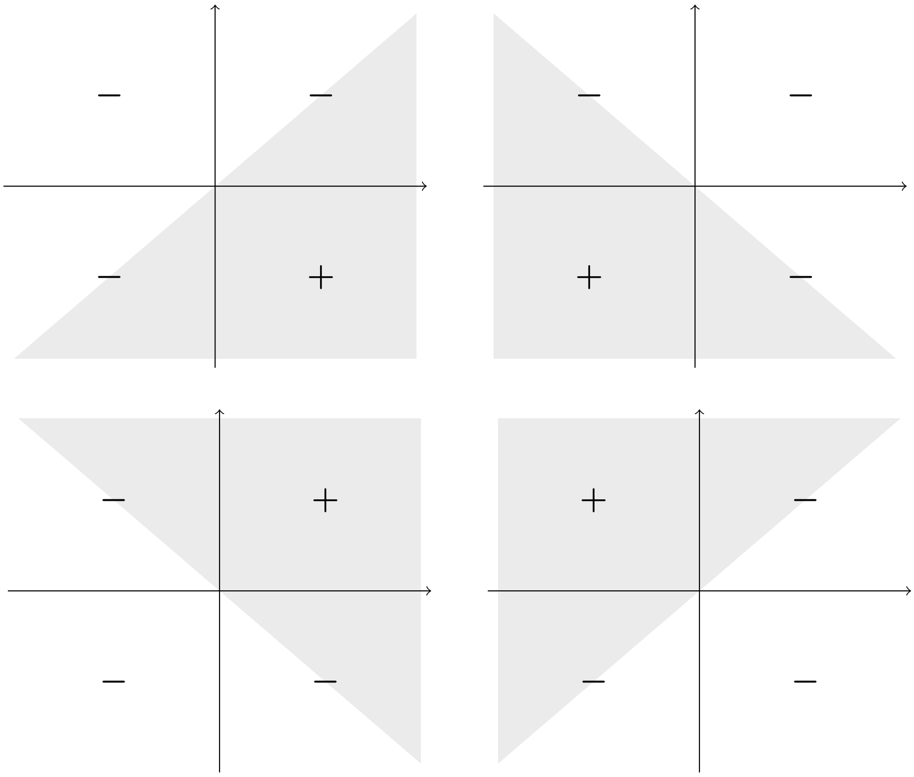


Figure 5.5: Using four fixed points, one can find the singletons of size four as a subclass of the open, homogeneous half planes. A co-singleton class of size four is induced by the complementary, closed half planes.

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 35: clearly, $\mathcal{C}^-$ consists of all open and closed pie slices with apex $\mathbf{0}$ *and $\mathcal{C}^+$ consists of the complements of such pie slices (see Figure 5.3). It is also easy to see that both s^+ and s^- are at least four (see Figure 5.5). We can see that four is also an upper bound for 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.

5.3.2 About the Optimal Choice of $\mathcal{H}$

In light of Lemma 33 one should choose $\mathcal{H}$ in such a way to minimize $\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C}, \mathcal{H})$. We are interested in bounding the possible reduction by combinatorial parameters, i.e., we want to give an upper bound to

$$\frac{\mathrm{VCdim}(\mathcal{C}) \cdot \theta(\mathcal{C})}{\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C}, \mathcal{H})} \ . \tag{5.3}$$

To avoid trivial cases, we will only consider classes $\mathcal{C}$ with $|\mathcal{C}| \geq 2$, so that for any $\mathcal{H} \supseteq \mathcal{C}$ holds $\mathrm{VCdim}(\mathcal{C}) \geq 1$ and $\theta(\mathcal{C}, \mathcal{H}) \geq 1$.

We prove the following bound on $\theta(\mathcal{C})$:

Lemma 42. *For any class $\mathcal{C}$ holds the inequality $\theta(\mathcal{C}) \leq s^+(\mathcal{C}) + s^-(\mathcal{C}) \leq s(\mathcal{C})/2$.*

Proof. We only have to prove the first inequality (the second one is trivial). For any $X' \subseteq X$ and $c^* \in \mathcal{C}$ let DIS^- be the part of the disagreement region that lies outside of the target concept c^*:

$$DIS^-(V_{\mathcal{C}}(X', c^*)) := \{x \in X : c^*(x) = 0, \exists h \in V_{\mathcal{C}}(X', c^*) \text{ s.t. } h(x) = 1\}$$

We want to cover this set with hypothesis from the version space. Note that the probability mass of the largest intersection of a hypothesis with DIS^- is at most $\sup_{h \in V_{\mathcal{C}}(X', c^*)} P(h \neq c^*)$ for an arbitrary domain distribution P. Thus every cover of DIS^- must at least use the following number of hypotheses:

$$\theta^+ := \frac{P(DIS^-(V_{\mathcal{C}}(X', c^*)))}{\sup_{h \in V_{\mathcal{C}}(X', c^*)} P(h \neq c^*)}$$

Since each element in a minimal cover contains at least one point which is not covered by any other element, we can conclude that $\mathcal{C}$ contains a singleton subclass of size θ^+, i.e., $\theta^+ \leq s^+(\mathcal{C})$. Analogously, $\mathcal{C}$ contains a co-singleton subclass of size θ^-, with a suitably defined θ^-. The result follows from $\theta(\mathcal{C}) = \sup_{P, X', c^*}(\theta^+ + \theta^-)$. □

We can now upper bound the ratio (5.3) easily:

$$\underbrace{\frac{\mathrm{VCdim}(\mathcal{C})}{\mathrm{VCdim}(\mathcal{H})}}_{\leq 1} \cdot \frac{\theta(\mathcal{C})}{\theta(\mathcal{C}, \mathcal{H})} \leq \frac{s(\mathcal{C})/2}{1}$$

Thus, the the maximal reduction achievable by choosing $\mathcal{H}$ well is of order $s(\mathcal{C})$. This bound is tight in the sense as there are classes that realize this ratio up to a constant factor:

Example 43. *Let X be the set $\{1,\ldots,2n\}$ and let $\mathcal{C}$ be the following combination of a power set and singletons:*

$$\mathcal{C} = \{A \cup \{b\} : A \subseteq \{1,\ldots,n\}, b \in \{n+1,\ldots,2n\}\}$$

It is easily to see that $s(\mathcal{C}) = \Theta(n)$ and $\mathrm{VCdim}(\mathcal{C}) \cdot \theta(\mathcal{C}) = \Theta(n) \cdot \Theta(n) = \Theta(n^2)$, while $\mathrm{VCdim}(\mathcal{H}) \cdot \theta(\mathcal{C},\mathcal{H}) = \Theta(n) \cdot \Theta(1) = \Theta(n)$ with the choice $\mathcal{H} = 2^X$.

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

Proof. The probability measures P on $X^{[k]}$ can be written as convex combinations of probability measures on the $X^{(i)}$, i.e., $P = \lambda_1 P_1 + \cdots + \lambda_k P_k$ where P_i is a probability measure on $X^{(i)}$, and the λ_i are non-negative numbers that sum-up to 1. A sample set $S \subseteq X^{[k]}$ 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,

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

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 $P_i, S^{(i)}, c^{(i)}$ properly. □

5.3.4 Comparison to Hanneke's Disagreement Coefficient

Hanneke's original definition of the disagreement coefficient is as follows:

Definition 45 (Hanneke [Han07]). *For $r > 0$ and $c^* \in \mathcal{C}$, let $B(c^*, r)$ be the ball of radius r around concept c^*, i.e. $B(c^*, r) := \{c \in \mathcal{C} : P(c \neq c^*) \leq r\}$. Then* Hanneke's disagreement coefficient *is defined as*

$$\theta_H(\mathcal{C}|P, c^*) := \sup_{r>0} \frac{P(DIS(B(c^*, r)))}{r} \ .$$

Taking the supremum over all $c^* \in \mathcal{C}$ and all distributions P over X yields the distribution independent disagreement coefficient

$$\theta_H(\mathcal{C}) := \sup_{P, c^*} \theta_H(\mathcal{C}|P, c^*) \ .$$

We would like to compare our variant of the disagreement coefficient with Hanneke's definition:

Lemma 46. *For any class $\mathcal{C}$, any $c^* \in \mathcal{C}$, any $X' \subseteq X$ and any distribution P holds*

$$\theta(\mathcal{C}|P, X', c^*) \leq \theta_H(\mathcal{C}|P, c^*) \ .$$

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

$$\begin{aligned}\theta(\mathcal{C}|P, X', c^*) &= \frac{P(DIS(V_{\mathcal{C}}(X', c^*)))}{r^*} \\ &\leq \frac{P(DIS(B(c^*, r^*)))}{r^*} \\ &\leq \theta_H(\mathcal{C}|P, c^*) \ .\end{aligned}$$

□

The following two observations demonstrate that our disagreement coefficient can indeed be much smaller than Hanneke's.

Lemma 47. *For any class $\mathcal{C}$ over X it holds that* $\mathrm{VCdim}(\mathcal{C}) \leq \theta_H(\mathcal{C}) \leq |X|$.

Proof. The first inequality is trivial for classes of VC-dimension zero, since θ_H is non-negative. Now let $X' \subseteq X$ be a set of size $d \geq 1$ shattered by $\mathcal{C}$ and let P be the uniform distribution over X'. Then $\theta_H(\mathcal{C}|c^*, P) \geq d$ holds for all $c^* \in \mathcal{C}$: let $r^* = 1/d$, so that $X' \subseteq DIS(B(c^*, r^*))$. Since $P(X') = 1$, it follows that $\theta_H(\mathcal{C}|P, c^*) \geq 1/(1/d) = d$.

The second inequality is trivial for infinite X. Note that $\theta_H(\mathcal{C}) \leq |X|$ holds for finite X, because $DIS(B(c^*, r)) \subseteq \{x \in X : P(\{x\}) \leq r\}$. □

Lemma 48. *Let $\mathcal{C}$ be a concept class, $c^* \in \mathcal{C}$, P a domain distribution and $X' \subseteq X$. If a concept $\tilde{c} \in V_{\mathcal{C}}(X', c^*)$ exists, that errs on the whole disagreement region almost surely under P, then $\theta(\mathcal{C}|P, X', c^*) = 1$.*

Proof. From $\sup_{c \in V_{\mathcal{C}}(X',c^*)} P(c \neq c^*) \geq P(\tilde{c} \neq c^*) \geq P(DIS(V_{\mathcal{C}}(X', c^*)))$ follows that $\theta(\mathcal{C}|P, X', c^*) \leq 1$. On the other hand, the error of a concept from the version space can never be larger than the probability mass of the disagreement region. Thus it holds that $\theta(\mathcal{C}|P, X', c^*) \geq 1$. □

Example 49. *Let $X \neq \emptyset$ be a finite set and let $\mathcal{C} = 2^X$. From Lemmas 47 and 48 follows $\theta_H(\mathcal{C}) = |X| \geq 1 = \theta(\mathcal{C})$.*

5.3.5 A Remark on PAC-learning Intersection-closed Classes

In this section we will shortly set our goal of analyzing supervised Co-training aside and prove a tight upper bound on the sample complexity of learning intersection-closed classes in the PAC-learning framework without extra assumptions. This answers Warmuth's conjecture [War04] mentioned in Section 5.1.2 in the positive.

While the main application of Hanneke's disagreement coefficient θ_H is in the field of agnostic active learning, Hanneke also proved the following theorem for the passive, realizable framework:

Theorem 50 (Hanneke [Han09]). *Let $\mathcal{C}$ be a concept class of VC-dimension d, P a distribution over X, $\boldsymbol{x} \sim P^m$ a random sample of size m and $c^* \in \mathcal{C}$. Then for any $\delta > 0$ it holds with probability $\geq 1 - \delta$ for all $h \in V_{\mathcal{C}}(\boldsymbol{x}, c^*)$*

$$P(h \neq c^*) \leq \frac{24}{m}\left(d \ln(880\, \theta_H(\mathcal{C}|P, c^*)) + \ln\frac{12}{\delta}\right) .$$

Thus the sample complexity $m^*_{\mathcal{C}}$ is upper bounded by $O\left(\frac{1}{\varepsilon}(d \ln \theta_H(\mathcal{C}) + \ln\frac{1}{\delta})\right)$. This bound is realized by any consistent, proper learning algorithm.

Combining Hanneke's result with our definition of the disagreement coefficient yields the following improved theorem, where we can replace θ_H by θ:

Theorem 51. *Let $\mathcal{C}$ be a concept class of VC-dimension d, P a distribution over X, $\boldsymbol{x} \sim P^m$ a random sample of size m and $c^* \in \mathcal{C}$. Then for any $\delta > 0$ it holds with probability $\geq 1 - \delta$ for all $h \in V_{\mathcal{C}}(\boldsymbol{x}, c^*)$*

$$P(h \neq c^*) \leq \frac{24}{m}\left(d \ln(880\, \theta(\mathcal{C}|P, c^*)) + \ln\frac{12}{\delta}\right) .$$

Thus the sample complexity $m^*_{\mathcal{C}}$ is upper bounded by $O\left(\frac{1}{\varepsilon}(d \ln \theta(\mathcal{C}) + \ln\frac{1}{\delta})\right)$. This bound is realized by any consistent, proper learning algorithm.

Proof. By carefully reading the proof of Theorem 50 in [Han09, Theorem 2.23 on page 51f], one can see that $\theta_H(\mathcal{C}|P, c^*)$ may safely be replaced by $\theta(\mathcal{C}|P, c^*)$. Since the proof of Theorem 51 is therefore the same as Hanneke's proof for Theorem 50—except for this minor change—we will only provide a rough proof sketch for the sake of completeness:

We want to prove the following statement by induction on m: "For all $\delta > 0$ it holds with a probability of at least $1 - \delta$ over the draw of $\boldsymbol{x} \sim P^m$ that

$$\sup_{h \in V_{\mathcal{C}}(\boldsymbol{x}, c^*)} P(h \neq c^*) \leq \frac{24}{m}\left(d \ln(880\, \theta(\mathcal{C}|P, c^*)) + \ln \frac{12}{\delta}\right) ."$$

For $m \leq d$ the statement is obvious, since the error is upper bounded by 1 (and $\theta(\mathcal{C}|P, c^*) \geq 1$ always). For the inductive step from $m/2$ to m we denote by V_m and $V_{m/2}$ the version spaces after drawing m or $m/2$ points of the sample, respectively. Note that $V_m \subseteq V_{m/2}$ and therefore $\sup_{h \in V_m} P(h \neq c^*) \leq \sup_{h \in V_{m/2}} P(h \neq c^*)$.

Let $DIS := DIS(V_{m/2})$. If $P(DIS) \leq \frac{8}{m} \cdot \ln \frac{3}{\delta}$, we are finished. In the other case a Chernoff-bound shows that at least $n := m/4 \cdot P(DIS)$ sample points from the second $m/2$ draws lie in DIS with a probability of at least $1 - \delta/3$. We can regard these n points as a random sample drawn according to P conditioned on DIS and apply a standard PAC bound from Blumer et al. [BEHW89] (in fact a step in the proof of Theorem 4), showing that $\sup_{h \in V_m} P(h \neq c^*)$ is upper bounded by $P(DIS) \cdot \frac{4}{n}\left(d \ln \frac{2en}{d} + \ln \frac{12}{\delta}\right)$ with a probability of at least $1 - \delta/3$. To handle the occurrence of n inside the logarithm, we note that $n \leq m/4 \cdot \theta(\mathcal{C}|P, c^*) \cdot \sup_{h \in V_{m/2}} P(h \neq c^*)$ by the definition of θ, i.e., Definition 30. After using the inductive assumption (with the probability of success set to $1 - \delta/3$) and some simplifying we arrive at the desired statement. □

We can now give the main result of this section:

Theorem 52. *Let $\mathcal{C}$ be an intersection-closed concept class of VC-dimension d. Then the sample complexity of $\mathcal{C}$ is upper bounded by*

$$m^*_{\mathcal{C}} = O\left(\frac{d + \ln(1/\delta)}{\varepsilon}\right) ,$$

which matches the general lower bound and is therefore optimal. The closure algorithm realizes this bound.

Proof. The idea of the proof is similar to that of the inductive step from the proof of Theorems 50 and 51.

Let $c^* \in \mathcal{C}$ and P be a distribution over X and let $\boldsymbol{x} \sim P^m$ be a random sample of size $m > 0$. By $h := \cap_{c \in V_{\mathcal{C}}(\boldsymbol{x}, c^*)} c$ we denote the closure algorithm's hypothesis after seeing the sample $(\boldsymbol{x}, c^*(\boldsymbol{x}))$. We will prove that $P(h \neq c^*) \leq \frac{48}{m}(d \ln 880 + \ln \frac{24}{\delta})$ holds with a probability of at least $1 - \delta$ over the draw of the sample.

Partition X into $X^+ := c^*$ and $X^- := X \setminus c^*$. Note that the cases $P(X^+) = 0$ and $P(X^-) = 0$ are trivial: in the former case $P(h \neq c^*)$ is zero and in the latter case we can skip a part of the proof and directly apply Theorem 51 and Lemma 48. Therefore we will assume $P(X^+), P(X^-) > 0$ and let P^+ and P^- denote P conditioned on X^+ and X^-. That is, for any $X' \subseteq X$ we define:

$$P^+(X') := P(X' \cap X^+)/P(X^+)$$
$$P^-(X') := P(X' \cap X^-)/P(X^-)$$

Since h only errs on positively labeled points it holds that

$$P(h \neq c^*) = P^+(h \neq c^*) \cdot P(X^+) + \underbrace{P^-(h \neq c^*)}_{=0} \cdot P(X^-) \ . \tag{5.4}$$

Thus, if $P(X^+) \leq \frac{48}{m}(d \ln 880 + \ln \frac{24}{\delta})$ we are already finished. So assume the opposite.

Let m^+ denote the number of elements in the sample that lie in X^+. Obviously, m^+ is binomially distributed with an expected value of $m \cdot P(X^+)$. By our assumption $P(X^+)$ is larger than $\frac{48}{m}(d \ln 880 + \ln \frac{24}{\delta}) \geq \frac{8}{m} \cdot \ln \frac{2}{\delta}$. From a Chernoff-bound follows that

$$m^+ \geq \frac{m}{2} \cdot P(X^+) \tag{5.5}$$

holds with a probability of at least $1 - \delta/2$.

The hypothesis class $\mathcal{H}$ of the closure algorithm is the class $\mathcal{C}^-$ from Definition 38, which contains *all* intersections, even over infinitely many concepts from $\mathcal{C}$. As Auer and Ortner have shown in [AO07], for intersection-closed classes $\mathcal{C}$ holds $\mathrm{VCdim}(\mathcal{C}^-) = \mathrm{VCdim}(\mathcal{C}) = d$. Thus, by Theorem 51 it holds with a probability of at least $1 - \delta/2$ that

$$P^+(h \neq c^*) \leq \frac{24}{m^+}\left(d \ln(880\, \theta(\mathcal{C}|P^+, c^*)) + \ln \frac{24}{\delta}\right) \ . \tag{5.6}$$

Since h errs on the whole disagreement region under P^+ for any sample, Lemma 48 yields $\theta(\mathcal{C}|c^*, P^+) = 1$.

We now obtain the desired result by plugging (5.6) and (5.5) in (5.4) and observing that the overall probability of failure is at most $\delta/2 + \delta/2 = \delta$. □

We briefly note that a corresponding theorem holds for union-closed classes (by taking the complement of each concept over X).

5.4 Supervised Learning and Co-training

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

$$P(h \neq c^*) = p_+ \cdot P(h \neq c^*|+) + p_- \cdot P(h \neq c^*|-) \ . \tag{5.7}$$

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 (5.7), 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 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.

Let us first fix some more notation that is 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 $DIS_1 = DIS(V_1)$, $DIS_2 = DIS(V_2)$ 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.

5.4.1 General Upper Bounds on the Sample Size

Three resolution 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 c^*.
2. Conditioned to "$+$", (x_1, x_2) is chosen at random according to $P(\cdot|+) \times P(\cdot|+)$. Conditioned to "$-$", (x_1, x_2) is chosen at random according to $P(\cdot|-) \times P(\cdot|-)$.

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 $x_1 \notin DIS_1$ or $x_2 \notin DIS_2$. As for the case $x_1 \in DIS_1$ and $x_2 \in DIS_2$, 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 53. *The number of labeled examples sufficient for learning $(\mathcal{C}_1, \mathcal{C}_2)$ in the PAC Co-training Model under the Conditional Independence Assumption by learners applying one of the rules R1, R2, R3 is given asymptotically as follows:*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \frac{\theta_{min}}{p_{min}}}\right) & \text{if rule R1 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \max\left\{\frac{1}{p_{min}}, \theta_{max}\right\}}\right) & \text{if rule R2 is applied} \\ \tilde{O}\left(\sqrt{\frac{d_1 d_2}{\varepsilon} \cdot \theta_1 \theta_2}\right) & \text{if rule R3 is applied} \end{cases} \tag{5.8}$$

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 $x_1 \in DIS_1$ and $x_2 \in DIS_2$.

Assume first that ambiguities are resolved according to rule R1. Note that the sample size specified in (5.8) is, by Theorem 4, 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 $x_2 \in DIS_2$. Thus the error rate induced by R1 is bounded (with high probability) as follows:

$$\begin{aligned}
&P(h_1(x_1) = 0 \wedge x_2 \in DIS_2|+)p_+ + P(h_1(x_1) = 1 \wedge x_2 \in DIS_2|-)p_- \\
\leq\ & \frac{1}{p_{min}} \cdot \Big(P(h_1(x_1) = 0|+)p_+ \cdot P(x_2 \in DIS_2|+)p_+ \\
&\qquad + P(h_1(x_1) = 1|-)p_- \cdot P(x_2 \in DIS_2|-)p_-\Big) \\
\leq\ & \frac{1}{p_{min}} \cdot \underbrace{\Big(P(h_1(x_1) = 0|+)p_+ + P(h_1(x_1) = 1|-)p_-\Big)}_{\leq \varepsilon_1} \\
&\qquad \cdot \underbrace{\Big(P(x_2 \in DIS_2|+)p_+ + P(x_2 \in DIS_2|-)p_-\Big)}_{\leq \theta_2\varepsilon_2 = \theta_{min}\varepsilon_2} \\
\leq\ & \frac{\theta_{min}}{p_{min}} \cdot \varepsilon_1\varepsilon_2 \\
=\ & \varepsilon
\end{aligned}$$

The first inequality in this calculation makes use of Conditional Independence and the third applies Lemma 33.

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

$$\begin{aligned}
&(h_1(x_1) = h_2(x_2) = 0 \wedge c^*(x_1, x_2) = 1) \\
\vee\ &(x_1 \in DIS_1 \wedge h_2(x_2) = 1 \wedge c^*(x_1, x_2) = 0) \\
\vee\ &(x_2 \in DIS_2 \wedge h_1(x_1) = 1 \wedge c^*(x_1, x_2) = 0)\ .
\end{aligned}$$

Thus, the error rate induced by R2 is bounded (with high probability) as follows:

$$\begin{aligned}
& P(h_1(x_1) = h_2(x_2) = 0|+)p_+ + P(x_1 \in DIS_1 \wedge h_2(x_2) = 1|-)p_- \\
& + P(x_2 \in DIS_2 \wedge h_1(x_1) = 1|-)p_- \\
\leq \quad & \frac{1}{p_+} \cdot \underbrace{P(h_1(x_1) = 0|+)p_+}_{\leq \varepsilon_1} \cdot \underbrace{P(h_2(x_2) = 0|+)p_+}_{\leq \varepsilon_2} \\
& + \underbrace{\tfrac{1}{p_-}}_{\leq 4} \cdot \Big(\underbrace{P(x_1 \in DIS_1|-)p_-}_{\leq \theta_1\varepsilon_1} \cdot \underbrace{P(h_2(x_2) = 1|-)p_-}_{\leq \varepsilon_2} \\
& \qquad + \underbrace{P(x_2 \in DIS_2|-)p_-}_{\leq \theta_2\varepsilon_2} \cdot \underbrace{P(h_1(x_1) = 1|-)p_-}_{\leq \varepsilon_1} \Big) \\
\leq\; & \frac{1}{p_{min}} \cdot \varepsilon_1\varepsilon_2 + 4\varepsilon_1\varepsilon_2 \cdot (\theta_1 + \theta_2) \\
\leq\; & \left(\frac{1}{p_{min}} + 8\theta_{max} \right) \cdot \varepsilon_1\varepsilon_2 \\
\leq\; & 2 \cdot \max \left\{ \frac{1}{p_{min}}, 8\theta_{max} \right\} \cdot \varepsilon_1\varepsilon_2 \\
\leq\; & \varepsilon
\end{aligned}$$

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 $x_1 \in DIS_1$, $x_2 \in DIS_2$, and the true label is "–". Thus the error rate induced by R3 is bounded (with high probability) as follows:

$$\begin{aligned}
& P(x_1 \in DIS_1 \wedge x_2 \in DIS_2|-)p_- \\
= \; & \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{P(x_1 \in DIS_1|-)p_-}_{\leq \theta_1\varepsilon_1} \cdot \underbrace{P(x_2 \in DIS_2|-)p_-}_{\leq \theta_2\varepsilon_2} \\
\leq\; & 4\theta_1\theta_2\varepsilon_1\varepsilon_2 \\
\leq\; & \varepsilon
\end{aligned}$$

This concludes the proof of the theorem. □

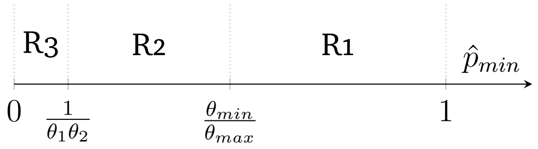

R3
R2
R1
$\hat{p}_{min}$
0
$\frac{1}{\theta_1\theta_2}$
$\frac{\theta_{min}}{\theta_{max}}$
1

Figure 5.6: The resolution rules used by the Combined Rule in dependence on $\hat{p}_{min}$ according to (5.9).

The Combined Rule

We now describe a strategy named "Combined Rule" that uses rules R1, R2, R3 as sub-routines. Given $(x_1, x_2) \in DIS_1 \times DIS_2$, 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

$$\begin{cases} \text{R1} & \text{if } \frac{\theta_{min}}{\theta_{max}} \leq \hat{p}_{min} \\ \text{R2} & \text{if } \frac{1}{\theta_1\theta_2} \leq \hat{p}_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \text{R3} & \text{if } \hat{p}_{min} < \frac{1}{\theta_1\theta_2} \end{cases} . \tag{5.9}$$

See Figure 5.6 for a visualization of these cases.

Theorem 54. *If the learner applies the Combined Rule, then*

$$\begin{cases} \tilde{O}\left(\sqrt{\frac{d_1d_2}{\varepsilon}\cdot\frac{\theta_{min}}{p_{min}}}\right) & \textit{if } \frac{\theta_{min}}{\theta_{max}} \leq p_{min} \\ \tilde{O}\left(\sqrt{\frac{d_1d_2}{\varepsilon}\cdot\theta_{max}}\right) & \textit{if } \frac{1}{\theta_{max}} \leq p_{min} < \frac{\theta_{min}}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1d_2}{\varepsilon}\cdot\frac{1}{p_{min}}}\right) & \textit{if } \frac{1}{\theta_1\theta_2} \leq p_{min} < \frac{1}{\theta_{max}} \\ \tilde{O}\left(\sqrt{\frac{d_1d_2}{\varepsilon}\cdot\theta_1\theta_2}\right) & \textit{if } p_{min} < \frac{1}{\theta_1\theta_2} \end{cases} \tag{5.10}$$

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 (5.10), we have at least

$$\tilde{O}\left(\sqrt{\frac{\min\{\theta_1\theta_2, 1/p_{min}\}}{\varepsilon}}\right) \tag{5.11}$$

labeled examples at our disposal. Furthermore, as illustrated in Figure 5.7, the upper bound from (5.10) is a continuous function in p_{min} (even for $p_{min} = \theta_{min}/\theta_{max}, 1/\theta_{max}, 1/(\theta_1\theta_2)$).

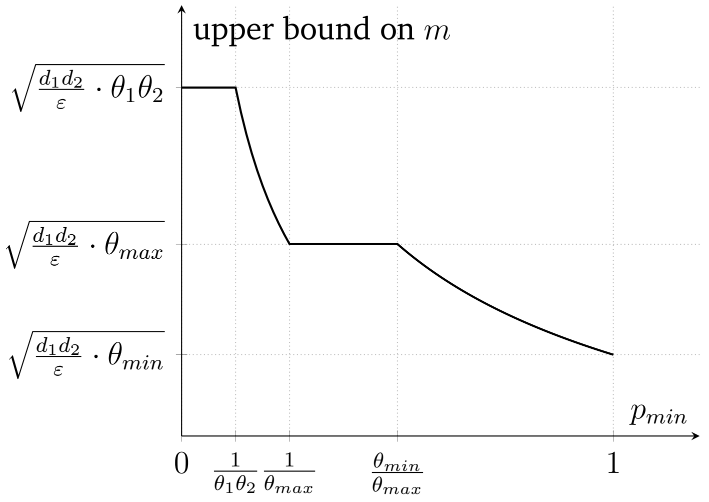


Figure 5.7: The sample complexity upper bound from (5.10) for the Combined Rule.

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 (5.11) 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 $p_{min} \le \varepsilon/4$.

Case 2: $2\theta_1\theta_2 \le \min\{4/\varepsilon, 1/p_{min}\}$ so that $p_{min} \le 1/(2\theta_1\theta_2)$ and $\varepsilon \le 2/(\theta_1\theta_2)$. Then, (5.11) is at least $\tilde{O}(\theta_1\theta_2)$. We can apply Chernoff-bounds and conclude that, with high probability, $\hat{p}_{min} \le \frac{1}{\theta_1\theta_2}$. But then rule R3 is applied which, according to Theorem 53, leads to the desired upper bound on the sample size.

Case 3: $1/p_{min} \le \min\{4/\varepsilon, 2\theta_1\theta_2\}$.
Now (5.11) is at least $\tilde{O}(1/p_{min})$. 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 $\hat{p}_{min} \ge \varepsilon/2$ so that the Combined Rule proceeds according to (5.9). If the learner could substitute the (unknown) p_{min} for $\hat{p}_{min}$ within (5.9), we could apply Theorem 53 and would be done. But since, as mentioned above, (5.10) is a *continuous* function in p_{min}, even the knowledge of $\hat{p}_{min}$ is sufficient. □

5.4.2 Lower Bounds on the Sample Size

A lower bound archetype

In this section, we prove a lower bound on the sample complexity for the class SF_n from Lemma 32. 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 chapter apply Lemma 55 directly or use the same proof technique.

Lemma 55. *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,

$$p_+ = p_{min} \leq \frac{1}{(n_1+2)(n_2+2)} = \frac{1}{\theta_1\theta_2} \ . \tag{5.12}$$

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 $c_1^* = \{a_0, a_s\} \in \mathcal{C}_1$. Similarly, index t is uniformly chosen at random from $X_2 \setminus \{0, 1\}$ and represents a randomly chosen target concept $c_2^* = \{b_0, b_t\} \in \mathcal{C}_2$. We define $P(a_s|+) = \sqrt{\varepsilon_1/p_+}$ and $P(b_t|+) = \sqrt{\varepsilon_2/p_+}$. 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.

- $P(a_0|+) = 1 - P(a_s|+)$ and $P(b_0|+) = 1 - P(b_t|+)$. 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.

- $P(a_1|-) = 1 - 4 \cdot \sqrt{n_1\varepsilon_1/n_2}$ and $P(b_1|-) = 1 - 4 \cdot \sqrt{n_2\varepsilon_2/n_1}$. 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 $1 - P(a_1|-)$. The instances from $X_2 \setminus \{b_0, b_1, b_t\}$ evenly share a minus-conditional probability mass of $1 - P(b_1|-)$.

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 $Z_b > n_b/2$ is at most $p_b = 1/5$.

- Using Equation (5.12) the expected 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, a_s does not occur in the sample with probability $p_3 = 1/40$. 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, we assume, to the advantage of the learner, that not only the labeled sample is revealed but also the probabilities p_+, $P(a_0|+)$, $P(b_0|+)$, $P(H_1|-)$, $P(H_2|-)$, 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 à-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, $P(a_s|+) = 1 - P(a_0|+) > 0$ whereas $P(x_1|+) = 0$ 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 à-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 à-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 à-posteriori probability given (x_1, x_2).
- Always vote for the label with the higher à-posteriori probability given U (the information that the test-instance belongs to U).
- If $P(E_+ \cap U) \geq P(E_- \cap U)$ vote always for "+", otherwise vote always for "−".

Clearly, the resulting Bayes-error equals $\min\{P(E_+ \cap U), P(E_- \cap U)\}$. It can be bounded from below as follows:

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

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

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

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 55 to classes with arbitrary VC-dimensions:

Corollary 56. *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 44.

A malicious scenario for the classes $\mathcal{C}_b^{[d_b]}$ is obtained by installing the malicious scenario from the proof of Lemma 55 (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 $i = 1, \dots, d_1$. Here, $s(i)$ is uniformly chosen at random from $\{2, \dots, n_1 + 2\}$. 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 $X_1^{(i)}$, $i = 1, \dots, d_1$, 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 $X_2^{(j)}$, $j = 1, \dots, d_2$.

Let us assume that the sample size satisfies
$$m \le \frac{\sqrt{d_1 d_2 n_1 n_2}}{40} \cdot \sqrt{1/\varepsilon}$$

In comparison to Lemma 55, 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 55) 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,
$$\mathbb{E}[Z_1^{(i)}] = \frac{1}{\sqrt{d_1 d_2}}(1 - p_+) \cdot 4 \cdot \sqrt{\frac{n_1 \varepsilon}{n_2}} \cdot m \le \frac{n_1}{10} \ .$$

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 55).

- The fact that $\mathbb{E}[Z_1^{(i)}]$ has the same upper bound as $\mathbb{E}[Z_1]$ is no accident: compared to Lemma 55, 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 55).

- 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 55). 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 55), 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 55, the Bayes-error is given by $\min\{P(E_+ \cap U), P(E_- \cap U)\}$, and an easy calculation (similar to the calculation in in the proof of Lemma 55) shows that it is bounded from below by $\varepsilon/4$.[5] □

[5] Factor 4 is lost in comparison to Lemma 55 because now half of the secret points in $X_1^{[d_1]}$ and $X_2^{[d_2]}$, respectively, might occur in the sample, whereas we could assume that a_s and b_t remained hidden from the learner in the proof of Lemma 55.

	$\Omega\left(\sqrt{\frac{1}{\epsilon}\cdot\frac{\theta_{min}}{p_{min}}}\right)$	$\Omega\left(\sqrt{\frac{1}{\epsilon}\cdot\theta_{max}}\right)$	$\Omega\left(\sqrt{\frac{1}{\epsilon}\cdot\frac{1}{p_{min}}}\right)$
$\mathcal{C}_1$	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	$\{a_0, \dots, a_{kn+2}\}$	$\{a_0, \dots, a_{kn+2}\}$	$\{a_0, \dots, a_{n+2}\}$
X_2	$\{b_0, \dots, b_{n+2}\}$	$\{b_0, \dots, b_{n+2}\}$	$\{b_0, \dots, b_{n+2}\}$
c_1^*	$\{a_0, a_s\}$	$\{a_0, a_s\}$	$\{a_0, a_s\}$
c_2^*	$X_2 \setminus \{b_0, b_t\}$	$X_2 \setminus \{b_0, b_t\}$	$\{b_0, b_t\}$
$P(a_s\|+)$	$2\sqrt{\frac{\epsilon}{np_+}}$	$2\sqrt{\frac{k\epsilon}{n}}$	$\sqrt{\frac{\epsilon}{p_+}}$
b_t	$P(b_t\|-) = 4\sqrt{\frac{\epsilon p_+}{n}}$	$P(b_t\|-) = 4\sqrt{\frac{\epsilon}{kn}}$	$P(b_t\|+) = \sqrt{\frac{\epsilon}{p_+}}$
$P(a_1\|-)$	$1 - \sqrt{\frac{n\epsilon}{p_+}}$	$1 - \sqrt{kn\epsilon}$	$1 - 8n \cdot \sqrt{\epsilon p_+}$
b_1	$P(b_1\|+) = 1 - \sqrt{\frac{n\epsilon}{p_+}}$	$P(b_1\|+) = 1 - \frac{1}{p_+}\sqrt{\frac{n\epsilon}{k}}$	$P(b_1\|-) = 1 - 8n \cdot \sqrt{\epsilon p_+}$
m_0	$\frac{1}{32}\sqrt{\frac{n}{\epsilon p_+}}$	$\frac{1}{32}\sqrt{\frac{kn}{\epsilon}}$	$\frac{1}{80}\sqrt{\frac{1}{\epsilon p_+}}$
$\mathbb{E}[Z_1]$	$p_- \cdot (1 - P(a_1\|-)) \cdot m$ $\leq \frac{kn}{32}$	$p_- \cdot (1 - P(a_1\|-)) \cdot m$ $\leq \frac{kn}{32}$	$p_- \cdot (1 - P(a_1\|-)) \cdot m$ $\leq \frac{n}{10}$
by Markov	$p_1 = P\left(Z_1 > \frac{kn}{2}\right) < \frac{1}{16}$	$p_1 = P\left(Z_1 > \frac{kn}{2}\right) < \frac{1}{16}$	$p_1 = P\left(Z_1 > \frac{n}{2}\right) < \frac{1}{5}$
$\mathbb{E}[Z_2]$	$p_+ \cdot (1 - P(b_1\|+)) \cdot m$ $\leq \frac{n}{32}$	$p_+ \cdot (1 - P(b_1\|+)) \cdot m$ $\leq \frac{n}{32}$	$p_- \cdot (1 - P(b_1\|-)) \cdot m$ $\leq \frac{n}{10}$
by Markov	$p_2 = P\left(Z_1 > \frac{n}{2}\right) < \frac{1}{16}$	$p_2 = P\left(Z_1 > \frac{n}{2}\right) < \frac{1}{16}$	$p_2 = P\left(Z_1 > \frac{n}{2}\right) < \frac{1}{5}$
$\mathbb{E}[\#a_s$ occurrence$]$	$p_+ \cdot P(a_s\|+) \cdot m \leq \frac{1}{16}$	$p_+ \cdot P(a_s\|+) \cdot m \leq \frac{kp_+}{16} \leq \frac{1}{8}$	$p_+ \cdot P(a_s\|+) \cdot m \leq \frac{1}{80}$
$P(a_s$ occurs$)$	$p_3 \leq \frac{1}{16}$	$p_3 \leq \frac{1}{8}$	$p_3 \leq \frac{1}{80}$
$\mathbb{E}[\#b_t$ occurrence$]$	$p_- \cdot P(b_t\|-) \cdot m \leq \frac{1}{8}$	$p_- \cdot P(b_t\|-) \cdot m \leq \frac{1}{8}$	$p_- \cdot P(b_t\|+) \cdot m \leq \frac{1}{80}$
$P(b_t$ occurs$)$	$p_4 \leq \frac{1}{8}$	$p_4 \leq \frac{1}{8}$	$p_4 \leq \frac{1}{80}$
$p_1 + p_2 + p_3 + p_4$	≤ 0.5	≤ 0.5	≤ 0.5
$P(U \cap E_+)$	$p_+ \cdot P(a_s\|+)\frac{1-P(b_1\|+)}{2}$ $\geq \epsilon$	$p_+ \cdot P(a_s\|+)\frac{1-P(b_1\|+)}{2}$ $\geq \epsilon$	$p_+ \cdot P(a_s\|+)P(b_t\|+)$ $\geq \epsilon$
$P(U \cap E_-)$	$p_- \cdot \frac{1-P(a_1\|-)}{2}P(b_t\|-)$ $\geq \epsilon$	$p_- \cdot \frac{1-P(a_1\|-)}{2}P(b_t\|-)$ $\geq \epsilon$	$p_- \cdot \frac{1-P(a_1\|-)}{2}\frac{1-P(b_1\|-)}{2}$ $\geq \epsilon$

Table 5.1: Overview of the necessary modifications to the proof of Lemma 55 to prove Theorem 57. Assume that $n, k \geq 2$.

Matching lower bounds for Theorem 54

Next we will provide lower bounds matching the upper bounds from the last section. Actually, Corollary 56 is a first result of this kind because the lower bound in this result matches with the upper bound in Theorem 54 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 Theorem 54 are fairly tight. The proof technique is the same as for the "archetypical" lower bounds.

For any concept class $\mathcal{C}$ over domain X, the class $\mathrm{co}(\mathcal{C})$ is given by

$$\mathrm{co}(\mathcal{C}) = \{X \setminus A \mid A \in \mathcal{C}\} \ .$$

Clearly, $\mathrm{VCdim}(\mathcal{C}) = \mathrm{VCdim}(\mathrm{co}(\mathcal{C}))$ and $\theta(\mathcal{C}) = \theta(\mathrm{co}(\mathcal{C}))$.

Theorem 57. *For sufficiently small $\varepsilon > 0$ at least*

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

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 56. Table 5.1 contains the necessary modifications to the proof of Lemma 55 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 56. □

Notice that the lower bounds in Theorem 57 nicely match with the general upper bounds given in Theorem 54.

5.4.3 Purely Combinatorial Bounds

In this section we will derive upper and lower bounds that are given in purely combinatorial parameters, i.e., without referring to disagreement coefficients or to p_{min}. Therefore, these bounds will be optimal in the worst case.

To this end we proceed as follows: first we give (almost) tight upper and lower bounds with succinct proofs, which unfortunately leave open the case of classes with (co-)singleton sizes smaller than three. We will close this gap in the final part of this section using a tedious case distinction rewarding us with (almost) tight bounds for all classes.

A combinatorial general upper bound

There is an alternative analysis for rule R3 which yields the following purely combinatorial upper bound on the sample size:

Theorem 58. *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 rule R3 is given asymptotically as:*

$$O\left(\sqrt{\frac{\max\{s_1^+ s_2^+, s_1^- s_2^-\}}{\varepsilon}}\right)$$

Proof. As in the proof of Theorem 53, we assume that $\hat{p}_{min} < 1/2$ holds if $p_{min} < 1/4$. And again, we may assume that $\theta_1 = \theta_{max}$ and $\hat{p}_- \geq 1/2$ so that $p_- \geq 1/4$.

From $\hat{p}_- \geq 1/2$ follows that R3 assigns label "+" to every instance $x_1 \in DIS_1$ (resp. to every instance $x_2 \in DIS_2$). 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, by Theorem 4, 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:

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

By symmetry, assuming that $\hat{p}_- < 1/2$, the upper bound $\tilde{O}\left(\sqrt{s_1^- s_2^- / \varepsilon}\right)$ 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 $\hat{p}_-$. □

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 [GG89], where it was shown that $\tilde{O}\left(s^-(\mathcal{C})/\varepsilon\right)$ many positive examples are sufficient (and necessary) to PAC-learn a class $\mathcal{C}$ from positive examples alone.

A combinatorial worst-case 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 upper bound from Theorem 58:

Theorem 59. *Assume that* $3 \leq s_b^+, s_b^- < \infty$ *for* $b = 1, 2$. *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 55, 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 55 with the special setting $p_+ = p_{min} = \varepsilon$ shows that at least $\Omega\Big(\sqrt{n_1 n_2/\varepsilon}\Big)$ many examples are needed for every pair $\mathcal{C}_1, \mathcal{C}_2$ of concept classes such that, for $b = 1, 2$, $\mathcal{C}_b$ contains a singleton subclass of size $n_b + 2$. □

Note that Theorem 59 implies a weak converse of Theorem 40 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 59 would exceed the upper bound for rule R3 in Theorem 58 with $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$.

Combinatorial bounds for classes with small (co-)singleton sizes

As noted in the introduction to this section, the matching combinatorial bounds from Theorem 58 and Theorem 59 do not hold for classes with (co-)singleton sizes smaller than three. We will show in the following that it is possible to drop this restriction and yield tight combinatorial bounds on the sample size that hold for all classes.

We will ignore the following two trivial cases: if $|\mathcal{C}_b| = 1$ or $|X_b| = 1$ for (at least) one b, then $m = 0$ or $m = 1$ examples are sufficient for learning. We will therefore assume in this section that $|X_b|, |\mathcal{C}_b| \geq 2$. Thus it also holds that $s_b^+, s_b^- > 0$.

First, we define two properties that characterize classes that are *not* trivial to learn: a class $\mathcal{C}$ over domain X has the property S^+ if it contains two concepts $c, \tilde{c}$ such that there are instances $x, \tilde{x} \in X$ with $c(x) = \tilde{c}(x) = 0$ and $c(\tilde{x}) \neq \tilde{c}(\tilde{x})$.

Analogously, a class $\mathcal{C}$ over domain X has the property S^- if it contains two concepts $c, \tilde{c}$ such that there are instances $x, \tilde{x} \in X$ with $c(x) = \tilde{c}(x) = 1$ and $c(\tilde{x}) \neq \tilde{c}(\tilde{x})$.

A few remarks about these properties are in order:

- Note that S^+ holds if s^+ is at least three (of course, the same is true for S^- and s^-). Furthermore, every class with $\mathrm{VCdim}(\mathcal{C}) \geq 2$ satisfies both properties.
- Since we assume $|X_b|, |\mathcal{C}_b| \geq 2$, we will always find an $\tilde{x} \in X_b$ and two concepts $c, \tilde{c} \in \mathcal{C}_b$ such that $c(\tilde{x}) \neq \tilde{c}(\tilde{x})$. If property S^+ is not satisfied, then c and $\tilde{c}$ must never agree on a negative label, i.e., for all $x \in X_b$ holds that $c(x) = 1$ or $\tilde{c}(x) = 1$. Note that c and $\tilde{c}$ may still agree on positive labels in this case.

We will show that the property S^+ in both C_1 and C_2 is necessary and sufficient for a term of $\sqrt{s_1^+ s_2^+ / \varepsilon}$ in the sample complexity bounds, while the property S^- is necessary and sufficient for a term of $\sqrt{s_1^- s_2^- / \varepsilon}$. Thus we have the following matching upper and lower bounds for the sample complexity (up to logarithmic factors):

	S^+ holds for both classes	otherwise
S^- holds for both classes	$\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}$	$\sqrt{s_1^- s_2^- / \varepsilon}$
otherwise	$\sqrt{s_1^+ s_2^+ / \varepsilon}$	1

Lemma 60. *If neither property S^+ nor property S^- is satisfied by both $\mathcal{C}_1$ and $\mathcal{C}_2$, then only one labeled example is sufficient and necessary to learn under the Conditional Independence Assumption.*

Proof. Note that if a class does not satisfy S^+ then one negatively labeled example is sufficient to learn the target concept exactly. The same holds for the dual with just one positive example. Therefore, if neither S^+ nor S^- is satisfied by both $\mathcal{C}_1$ and $\mathcal{C}_2$, then just one example is sufficient and, because of our assumption $|\mathcal{C}_b| > 1$, also necessary to learn exactly. □

We will now generalize the combinatorial bound from Theorem 59 to any pair of classes with property S^+ (resp. S^-):

Theorem 61. *Assume that $s_b^+ < \infty$ for $b = 1, 2$ and both $\mathcal{C}_b$ satisfy S^+. 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 $\Omega\Big(\sqrt{s_1^+ s_2^+ / \varepsilon}\Big)$.*

Proof. We consider three cases:

Case 1: For $s_1^+, s_2^+ \geq 3$ we apply Theorem 59.

Case 2: For $s_1^+ \geq 3$, $s_2^+ < 3$ it suffices to prove a bound of $m = \Omega\Big(\sqrt{s_1^+/\varepsilon}\Big)$. Let $\mathcal{C}_1$ be the class of singletons over $X_1 = \{0, \dots, n_1 + 1\}$ such that $n_1 = s_1^+ - 2 > 0$. Because property S^+ holds, we can identify $\mathcal{C}_2$ with the class $\{\emptyset, \{1\}\}$ over domain $X_2 = \{0, 1\}$.

Choose t_1 uniformly at random from $X_1 \setminus \{0\}$ and let the target for the first class be $c_1^* = \{t_1\}$. Choose the target c_2^* for the second class uniformly at random from $\{\emptyset, \{1\}\}$. Assume $\varepsilon < 1/(16n_1)$ and let the distribution P satisfy:

- $P(x_1 = 0|-) = 1 - 4\sqrt{n_1\varepsilon}$. The rest of the minus-conditional probability mass is evenly distributed over the remaining n_1 negative instances, i.e., $X_1 \setminus \{0, t_1\}$.
- If $c_2^* = \emptyset$, we set:
  $p_+ = 0$, $P(x_2 = 1|-) = 4\sqrt{\varepsilon/n_1}$ and $P(x_2 = 0|-) = 1 - 4\sqrt{\varepsilon/n_1}$.
- If $c_2^* = \{1\}$, we set:
  $p_+ = \varepsilon$ and $P(x_1 = t_1|+) = P(x_2 = 1|+) = P(x_2 = 0|-) = 1$.

Assume that $m \leq 1/40 \cdot \sqrt{n_1/\varepsilon}$. Now the following holds:

- The expected number of positive examples is 0 in the case $c_2^* = \emptyset$ and $p_+ \cdot m \leq 1/40 \cdot \sqrt{\varepsilon n_1} < 1/160$ in the case $c_2^* = \{1\}$. Therefore, regardless of the target concept, positive examples will only occur in the sample with a probability less than $1/160$.
- Let $X_1' \subseteq X_1 \setminus \{0, t_1\}$ denote the instances from the first class that occurred in the sample (these are the "interesting" negative examples for the first class). The expected number of negative examples, such that $x_1 \neq 0$ holds, is $p_- \cdot 4 \cdot \sqrt{\varepsilon n_1} \cdot m \leq n_1/10$. Thus, by the Markov-inequality, the probability that $|X_1'| \geq n_1/2$ holds is at most $1/5$.
- The expected number of negative examples such that $x_2 = 1$ (these are the "interesting" negative examples for the second class) is 0 in the case $c_2^* = \{1\}$ and $p_- \cdot 4 \cdot \sqrt{\varepsilon/n_1} \cdot m \leq 10$ in the case $c_2^* = \emptyset$. Thus interesting negative examples for the second class will only occur in the sample with a probability of at most $1/10$.

Consequently, with a probability of at least $1 - 1/160 - 1/5 - 1/10 > 1/2$ the learner will only see examples that have a negative label, are uninteresting for the second class and contain less than $n_1/2$ interesting examples for the first class. Note that the bound on the probability of this event is valid for any choice of c_2^*.

We will assume in the following that the event occurs. Note that the learner is unable to differentiate between the cases $p_+ = 0$ and $p_+ = \varepsilon$, i.e., she can not determine whether $c_2^* = \emptyset$ or $c_2^* = \{1\}$ holds. Furthermore, because of symmetry, the learner can only guess t_1 randomly out of $X_1 \setminus (X_1' \cup \{0\})$. It follows that the plus- and minus-conditional à-posteriori probabilities for pairs of the form $(x_1, 1)$, such that $x_1 \in X_1 \setminus (X_1' \cup \{0\})$, are equal. Thus the Bayes-decision will assign the same label to all such pairs, of which there are more than $n_1/2$. We do not have to derive the actual Bayes-decision, as it is easy to see that the learner will fail in any case:

If she prefers the negative label, there is a probability of $1/2$ that the target concept is $c_2^* = \{1\}$ and the error rate will be at least $p_+ = \varepsilon$. On the other hand, if she votes for the positive label, $c_2^* = \emptyset$ holds with a probability of $1/2$ and the error rate will be at least $p_- \cdot n_1/2 \cdot \frac{4\sqrt{n_1\varepsilon}}{n_1} \cdot 4\sqrt{\varepsilon/n_1} = 8\varepsilon \geq \varepsilon$ by conditional independence. Therefore, the Bayes-error is at least ε with a probability larger than $1/2 \cdot 1/2 = 1/4$.

Case 3: For the last case ($s_1^+, s_2^+ < 3$) it suffices to show a lower bound of $m = \Omega(\varepsilon^{-1/2})$. This can be done as follows:

Because the property S^+ holds for both classes, we can identify both $\mathcal{C}_1$ and $\mathcal{C}_2$ with the set $\{\emptyset, \{1\}\}$ over domain $X_1 = X_2 = \{0, 1\}$. We choose $c^* \in \{\emptyset, \{1\}\}$ uniformly at random and let $c_1^* = c_2^* = c^*$. Let P be the following distribution:

- If $c^* = \emptyset$, we set $p_+ = 0$, $P(x_1 = 1|-) = P(x_2 = 1|-) = \sqrt{\varepsilon}$ and $P(x_1 = 0|-) = P(x_2 = 0|-) = 1 - \sqrt{\varepsilon}$.
- If $c^* = \{1\}$, we set $p_+ = \varepsilon$, $P(x_1 = 1|+) = P(x_2 = 1|+) = 1$ and $P(x_1 = 0|-) = P(x_2 = 0|-) = 1$.

If we assume $m = o(\varepsilon^{-1/2})$, it is easy to see that with a high probability (for both choices of c^*) the sample contains the pair $(0, 0)$ only. Therefore, the learner is unaware of the true label of the pair $(1, 1)$.

If she decides to assign a positive label to the instance $(1, 1)$ and $c^* = \emptyset$ holds, the error rate is at least (by conditional independence) $p_- \cdot \sqrt{\varepsilon} \cdot \sqrt{\varepsilon} = \varepsilon$. On the other hand, if the learner chooses a negative label and $c^* = \{1\}$ holds, the error rate is at least $p_+ = \varepsilon$. Note that the probability of failure is at least $1/2$ in both cases.

Thus we need at least $\Omega(\varepsilon^{-1/2})$ many examples to learn successfully. □

Please note that the dual case "$s_b < \infty$ and properties S^- hold" leads to a dual bound with s_b^- instead of s_b^+. Thus, if both classes satisfy S^+ and S^-, we achieve the combinatorial lower bound $m = \Omega\Big(\sqrt{\max\{s_1^+ s_2^+, s_1^- s_2^-\}/\varepsilon}\Big)$ from Theorem 59 again.

We will now prove a matching upper bound. As a first step, we will have a short glance at one-sided errors (Section 5.4.4 provides a more detailed look at this topic):

Lemma 62. *Let $\mathcal{C}_b$ for $b \in \{1,2\}$ be two concept classes over X_b. Let resolution rule R$^+$ be the following: in the case $x_1 \in DIS_1$ and $x_2 \in DIS_2$ we return the label "$+$". Using resolution rule R$^+$,*

$$m = \tilde{O}\left(\sqrt{\frac{s_1^+ s_2^+}{\varepsilon \cdot p_-}}\right)$$

labeled examples are sufficient for PAC-learning.

Proof. Recall from Lemma 37 that the hypotheses classes $\mathcal{C}_b^+$, for $b = 1, 2$, are closed under union and thus contain hypotheses $h_b^\cup$ with minus-sided error, namely the unions of all hypotheses in the version spaces $V_{\mathcal{C}_b^+}$. Since the hypotheses $h_b^\cup$ are consistent and, by Lemma 37, $d_b := \mathrm{VCdim}(\mathcal{C}_b^+) \le s_b^+$, it follows from Theorem 4 that (with high probability) the error rates for $h_b^\cup$ after seeing at least $m = \tilde{O}\left(\sqrt{\frac{s_1^+ s_2^+}{\varepsilon \cdot p_-}}\right)$ examples are:

$$\varepsilon_1 = d_1 \cdot \sqrt{\frac{p_- \cdot \varepsilon}{s_1^+ s_2^+}} \le \sqrt{\frac{s_1^+ \cdot p_- \cdot \varepsilon}{s_2^+}}, \quad \varepsilon_2 \le \sqrt{\frac{s_2^+ \cdot p_- \cdot \varepsilon}{s_1^+}}$$

Because both $h_b^\cup$ have minus-sided error, they return a "+"-label on the whole disagreement region, thus they just behave like rule R$^+$.

Using conditional independence we can bound the error from above:

$$\begin{aligned}
&P(x_1 \in DIS_1 \wedge x_2 \in DIS_2|-) \cdot p_- \\
&= P(x_1 \in DIS_1|-) \cdot P(x_2 \in DIS_2|-) \cdot p_- \\
&= P\Big(h_1^\cup(x_1) = 1| - \Big) \cdot P\Big(h_2^\cup(x_2) = 1| - \Big) \cdot p_- \\
&\le \frac{1}{p_-} \cdot \underbrace{P\Big(h_1^\cup(x_1) = 1| - \Big)p_-}_{=\varepsilon_1} \cdot \underbrace{P\Big(h_2^\cup(x_2) = 1| - \Big)p_-}_{=\varepsilon_2} \\
&\le \varepsilon
\end{aligned}$$

□

A dual result (using $\mathcal{C}_b^-$ and $h_b^\cap$, which denotes the smallest consistent hypothesis) yields a bound of $m = \tilde{O}\Big(\sqrt{s_1^- s_2^- / (\varepsilon \cdot p_+)}\Big)$ for a correspondingly defined rule R$^-$, which only errs on positive points.

Now we can give a version of the upper bound from Theorem 58, which is (almost) tight for classes that do not satisfy S^+ or S^-:

Theorem 63. *The term "$s_1^+ s_2^+$" in the upper bound from Theorem 58 can be replaced by 1 if (at least) one of the classes does not satisfy S^+.*

The term "$s_1^- s_2^-$" in the upper bound from Theorem 58 can be replaced by 1 if (at least) one of the classes does not satisfy S^-.

Proof. We only show the result for the case "$\mathcal{C}_1$ does not satisfy S^-" (and both classes still satisfy S^+), but all other cases can be treated completely analogously. We now have to show that a sample size of

$$\tilde{O}\left(\sqrt{s_1^+ s_2^+ / \varepsilon}\right) \tag{5.13}$$

suffices for learning.

Recall from the proof of Lemma 60 that in the given setting just one example with a positive label is sufficient to learn the target exactly. A direct computation shows that a positive example will occur in a sample of size (5.13) with a high probability in the case $p_+ \geq \sqrt{\varepsilon/(s_1^+ s_2^+)}$. Thus we assume in the following that $p_+ < \sqrt{\varepsilon/(s_1^+ s_2^+)}$ holds.

If the learner's sample only contains positive examples, she proceeds using rule R^+. Since $1/p_- = 1/(1-p_+) = O(1)$ for small ε, the sample size (5.13) is large enough for an application of Lemma 62, which concludes the proof. □

5.4.4 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 64. *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 Theorem 54). 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 53, 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:

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

□

Note that the upper bound from Theorem 64 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 57.

Theorem 65. *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*

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

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. As in the proof of Theorem 64, 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 $x_1 \in DIS_1$ and $x_2 \in DIS_2$. Note that $x_1 \in DIS_1$ implies that $h_1(x_1) = 0$, and $x_2 \in DIS_2$ 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

[6] Recall from the proof of Theorem 54 that knowing the empirical estimate $\hat{p}_{min}$ instead of the true value p_{min} does not cause much trouble.

the empirically less likely label) otherwise. According to Theorem 53, 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 53, 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:

$$\begin{aligned} & P(x_1 \in DIS_1 \wedge h_2(x_2) = 1|-)p_- \\ \leq \; & \underbrace{\frac{1}{p_-}}_{\leq 4} \cdot \underbrace{P(h_1(x_1) = 0|-)p_-}_{\leq \theta_1 \varepsilon_1} \cdot \underbrace{P(h_2(x_2) = 1|-)p_-}_{\leq \varepsilon_2} \\ \leq \; & 4 \cdot \theta_{max} \cdot \varepsilon_1 \varepsilon_2 \\ = \; & \varepsilon \end{aligned}$$

□

Note that the upper bound from Theorem 65 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 57.

5.4.5 Generalization to k-tuples

In this section we give several bounds on the supervised sample complexity in a generalization of Co-training where the learner is provided with sample points consisting of k-tuples $(x_1, \ldots, x_k)$ instead of pairs. Also, we now have k concept classes $\mathcal{C}_1, \ldots \mathcal{C}_k$ over k domains $X_1, \ldots, X_k$. The assumptions of the original framework are adjusted in a straightforward way. Especially we again assume Conditional Independence given the label.

As we will see below, the bounds on the sample size do not change much compared to the old model, but note that the dependency on $1/\varepsilon$ is now in the order of the k-th root. The proofs are also very similar to the ones from the case $k = 2$.

We can give a generalization of the upper bound from Theorem 53. To this end we generalize the three resolution rules in the following way:

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

R2: If $h_1(x_1) = \ldots = h_k(x_k)$, then vote for the same label. Otherwise vote for the label that occurred less often in the sample.

R3: If $\hat{p}_- \geq 1/2$, then vote for label "$+$". Otherwise, vote for label "$-$" (this is the same rule as in the case $k = 2$).

Corollary 66. *The number of labeled examples sufficient for learning* $(\mathcal{C}_1, \ldots, \mathcal{C}_k)$ *in the PAC Co-training Model with* k*-tuples under the Conditional Independence Assumption by learners applying one of the rules R1, R2, R3 is given asymptotically as follows:*

$$\begin{cases} \tilde{O}\left(\left(\frac{d_1\cdots d_k}{\epsilon}\cdot\frac{1}{p_{min}^{k-1}}\cdot\frac{\theta_1\cdots\theta_k}{\theta_{max}}\right)^{1/k}\right) & \textit{if rule R1 is applied} \\ \tilde{O}\left(\left(\frac{d_1\cdots d_k}{\epsilon}\cdot\max\left\{\frac{1}{p_{min}^{k-1}}, \frac{\theta_1\cdots\theta_k}{\theta_{min}}\right\}\right)^{1/k}\right) & \textit{if rule R2 is applied} \\ \tilde{O}\left(\left(\frac{d_1\cdots d_k}{\epsilon}\cdot\theta_1\cdots\theta_k\right)^{1/k}\right) & \textit{if rule R3 is applied} \end{cases}$$

Proof. We follow the proof of Theorem 53 and only point out the necessary changes. We have to adjust the error rates ε_b properly and recalculate the overall error for each rule.

For R1, note that the sample size is sufficient to bound the error rates of the hypotheses h_b (with a high probability) by

$$\varepsilon_b = \left(\frac{d_b^k}{d_1\cdots d_k}\cdot p_{min}^{k-1}\cdot\frac{\theta_{max}}{\theta_1\cdots\theta_k}\cdot\varepsilon\right)^{1/k} .$$

We can now bound the error of R1 as follows (remember that we assume $\theta_1 = \theta_{max}$):

$$\begin{aligned} & P\Big(h_1(x_1) = 0 \wedge x_2 \in DIS_2 \wedge \ldots \wedge x_k \in DIS_k\Big|+\big)p_+ \\ & + P\Big(h_1(x_1) = 1 \wedge x_2 \in DIS_2 \wedge \ldots \wedge x_k \in DIS_k\Big|-\big)p_- \\ \leq\ & \frac{1}{p_{min}^{k-1}}\cdot\varepsilon_1\cdot\theta_2\varepsilon_2\cdots\theta_k\varepsilon_k \\ \leq\ & \frac{1}{p_{min}^{k-1}}\cdot\frac{\theta_1\cdots\theta_k}{\theta_{max}}\cdot\varepsilon_1\cdots\varepsilon_k \\ =\ & \varepsilon \end{aligned}$$

Since, obviously, $(k \cdot 4^{k-1})^{1/k} = O(1)$, we can bound the error rate of h_b for rule R2 by

$$\varepsilon_b = \left(\frac{d_b^k}{2d_1\cdots d_k}\cdot\min\left\{p_{min}^{k-1}, \frac{\theta_{\min}}{k\cdot 4^{k-1}\cdot\theta_1\cdots\theta_k}\right\}\cdot\varepsilon\right)^{1/k} .$$

This yields the following upper bound for the error of R2:

$$\begin{aligned}
&P\Big(h_1(x_1) = \ldots = h_k(x_k) = 0| + \Big)p_+ \\
&+ \sum_{i=1}^{k} P\Big(h_i(x_i) = 1 \wedge x_j \in DIS_j \ \forall j \neq i \Big| -)p_- \\
\leq\ & \frac{1}{p_+^{k-1}} \cdot \varepsilon_1 \cdots \varepsilon_k + \underbrace{\frac{1}{p_-^{k-1}}}_{\leq 4^{k-1}} \cdot \varepsilon_1 \cdots \varepsilon_k \cdot \underbrace{\sum_{j=1}^{k} \frac{\theta_1 \cdots \theta_k}{\theta_j}}_{\leq k \cdot \frac{\theta_1 \cdots \theta_k}{\theta_{min}}} \\
\leq\ & 2 \cdot \max\left\{\frac{1}{p_{min}^{k-1}}, k \cdot 4^{k-1} \cdot \frac{\theta_1 \cdots \theta_k}{\theta_{min}}\right\} \cdot \varepsilon_1 \cdots \varepsilon_k \\
\leq\ & \varepsilon
\end{aligned}$$

For R3, we can bound the error rate of h_b by

$$\varepsilon_b = \left(\frac{d_b^k}{d_1 \cdots d_k} \cdot \frac{\varepsilon}{4^{k-1} \cdot \theta_1 \cdots \theta_k}\right)^{1/k} .$$

Thus the error rate of R3 is given by:

$$\begin{aligned}
&P\Big(x_1 \in DIS_1 \wedge \ldots x_k \in DIS_k \Big| -)p_- \\
\leq\ & \frac{1}{p_-^{k-1}} \cdot \theta_1 \varepsilon_1 \cdots \theta_k \varepsilon_k \\
\leq\ & 4^{k-1} \cdot \theta_1 \cdots \theta_k \cdot \varepsilon_1 \cdots \varepsilon_k \\
=\ & \varepsilon
\end{aligned}$$

□

One can also generalize lower bounds in this model, as we will show here for the bound from Lemma 55:

Corollary 67. *Let $n_b \geq 1$, and let $\mathcal{C}_b = \mathrm{SF}_{n_b+2}$ so that $\theta_b = n_b + 2$ for $b = 1, \ldots, k$. Then, for every $p_{min} \leq 1/(\theta_1 \cdots \theta_k)^{1/(k-1)}$ and every sufficiently small $\varepsilon > 0$, the number of examples needed to learn $\mathcal{C}_1, \ldots, \mathcal{C}_k$ in the PAC Co-training Model with k-tuples under the Conditional Independence Assumption is at least*

$$\Omega\left(\frac{1}{k} \cdot \Big(n_1 \cdots n_k/\varepsilon\Big)^{1/k}\right) .$$

Proof. We proceed as in the proof of Lemma 55 and point out the necessary changes. Again, we assume that $p_{min} = p_+$ holds. Let $\mathcal{C}_b$ be a concept class over domain $X_b = \{a_0^b, \ldots, a_{n_b+1}^b\}$ for $b = 1, \ldots, k$. From $p_+ = p_{min} \leq 1/(\theta_1 \cdots \theta_k)^{1/(k-1)}$ follows that $(n_1 \cdots n_k)^{1/k} \cdot p_+^{\frac{k-1}{k}} \leq 1$. We change the malign scenario from Lemma 55 in the following way:

- Index s_b is chosen uniformly at random from $X_b \backslash \{0,1\}$ and let $P(a^b_{s_b}|+) = (\varepsilon/p_+)^{1/k}$.
- Let $P(a^b_0|+) = 1-(\varepsilon/p_+)^{1/k}$.
- Let $P(a^b_1|-) = 1-4\cdot\left(\frac{n_b^k}{n_1\cdots n_k}\cdot\varepsilon\right)^{1/k}$.
- The instances from $X_b \setminus \{a^b_0, a^b_1, a^b_{s_b}\}$ evenly share a minus-conditional probability mass of $4\cdot\left(\frac{n_b^k}{n_1\cdots n_k}\cdot\varepsilon\right)^{1/k}$.

Let us assume now that the sample size satisfies

$$m \le \frac{(n_1\cdots n_k)^{1/k}}{20\cdot k}\cdot\left(\frac{1}{\varepsilon}\right)^{1/k} \quad .$$

Reusing the notation Z_b (the number of "interesting" negative examples in the sample) from the proof of Lemma 55, we calculate:

- For the expectation of Z_b holds:

  $$\mathbb{E}[Z_b] \le p_-\cdot 4\cdot\left(\frac{n_b^k}{n_1\cdots n_k}\cdot\varepsilon\right)^{1/k}\cdot\frac{(n_1\cdots n_k)^{1/k}}{20\cdot k}\cdot\left(\frac{1}{\varepsilon}\right)^{1/k} \le \frac{n_b}{5\cdot k}$$

  Thus with a probability of at least $1-2/5$ at least half of the interesting negative points in each X_b remain hidden from the learner simultaneously.

- The expected number of occurrences of $a^b_{s_b}$ in the sample is bounded by

  $$\begin{aligned} & p_+\cdot\left(\frac{\varepsilon}{p_+}\right)^{1/k}\cdot\frac{(n_1\cdots n_k)^{1/k}}{20\cdot k}\cdot\left(\frac{1}{\varepsilon}\right)^{1/k} \\ \le\; & \frac{1}{20\cdot k}\cdot\underbrace{(n_1\cdots n_k)^{1/k}\cdot p_+^{\frac{k-1}{k}}}_{\le 1} \\ \le\; & \frac{1}{20\cdot k} \quad . \end{aligned}$$

  Thus with a probability of at least $1-1/20$ the learner never sees any $a^b_{s_b}$ in the sample.

Therefore, with a probability of at least $1-2/5-1/20 > 1/2$, the learner can only guess the label on instances that were not included in the sample. Analogously as in the proof of Lemma 55 we bound the Bayes-error, which is given by $\min\{P(U\cap E_+), P(U\cap E_-)\}$:

$$P(U\cap E_+) \ge p_+\cdot\prod_{b=1}^{k} P(a^b_{s_b}|+) = p_+\cdot\left(\left(\frac{\varepsilon}{p_+}\right)^{1/k}\right)^k = \varepsilon$$

$$P(U \cap E_-) \geq \underbrace{(1 - p_+)}_{\geq 1/2} \cdot \prod_{b=1}^{k} \frac{1}{2} \cdot 4 \cdot \left(\frac{n_b^k}{n_1 \cdots n_k} \cdot \varepsilon \right)^{1/k} \geq 2^{k-1} \cdot \varepsilon \geq \varepsilon$$

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

Furthermore, it is possible to apply the padding argument and generalize Corollary 56 to the k-tuple model in the same fashion. Note that the resulting lower bound for rule R3 is tight up to a previously unnoticeable factor of $1/k$ (and the usual logarithmic factors).

5.4.6 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 [BBY04]. 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 [BBY04] for this restriction stems from the fact that the assumption we have to make on 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 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 $c_1^* \in \mathcal{C}_1$ and $c_2^* \in \mathcal{C}_2$, meaning that $c_1^*(x_1) = c_2^*(x_2)$ with probability 1. Let P^+ denote the distribution conditioned on positive labels and let X_1^+ and X_2^+ denote the support of the marginal distributions of P^+ over X_1 resp. X_2. We say that P^+ is *α-expanding with respect to hypothesis classes H_1, H_2* if for all $h_1 \in H_1$ and $h_2 \in H_2$:

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

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

The authors of [BBY04] show that the α-expansion assumption is weaker than Conditional Independence and in [BB10] it is mentioned, that, if both P^+ and P^- 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 [BBY04]. 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 $c_1^* = c_2^* = c^*$ and non-empty. Note that this also implies $X_1^+ = X_2^+$. Let $p_+ = p_- = 1/2$. Distribution P^+ 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 $\{1, \dots, d\}$ 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 [BBY04].

Theorem 68. *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.

Proof. We will reduce learning the pair (c_1^*, c_2^*) to learning the rectangle c^* 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 $c_1^* = c_2^*$ follows that all information in the sample about c_1^* can also be applied to learn c_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 P^+. Let h be any rectangle contained in c^*. We denote the error of (h, h) with respect to the distribution P^+ by $P^+(h \neq c^*)$. To model the standard learning setting, let U be the uniform distribution over c^* and denote the error of h with respect to U by $U(h \neq c^*)$. We claim that

$$P^+(h \neq c^*) \geq \frac{d-1}{d} \cdot U(h \neq c^*) \ . \tag{5.14}$$

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

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

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 $U(h \neq c^*)$ 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 (5.14), 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 c^* 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

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

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 [BDLP08] 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 68, this is not possible since we fixed p_+.

Also observe that the bound from Theorem 68 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 64 that $\tilde{O}\left(d/\sqrt{\varepsilon}\right)$ labeled examples are sufficient for learning d-dimensional axis-aligned rectangles in the PAC Co-training Model under the Conditional Independence Assumption.

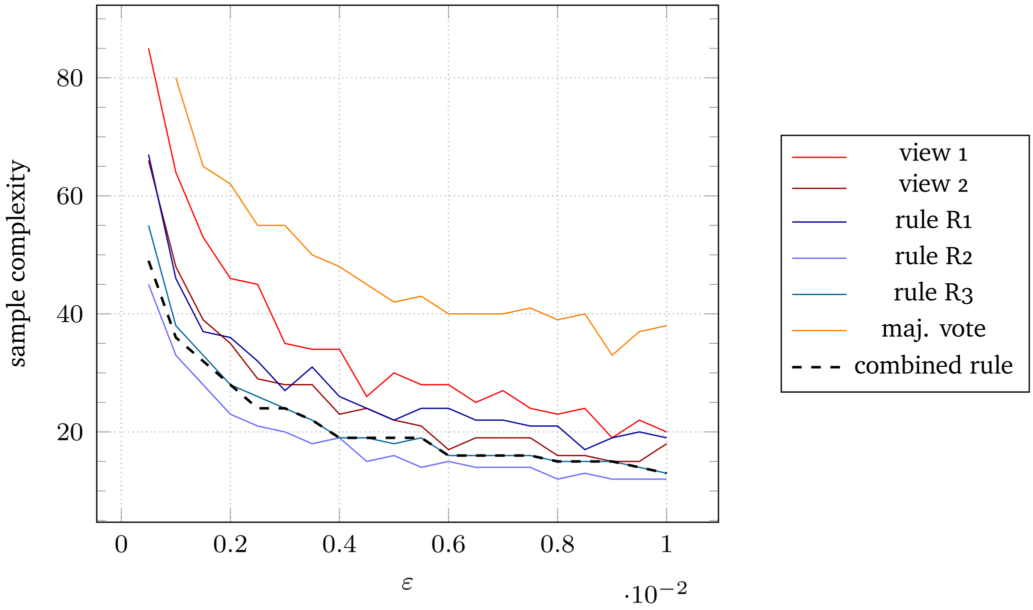


Figure 5.8: Estimated sample complexity under the worst-case distribution for $(\mathcal{C}_1, \mathcal{C}_2) = (\mathrm{SF}_{10}, \mathrm{SF}_5)$.

5.5 Experimental Results

In this section, we study the lower bound on the sample complexity proven in Lemma 55 by conducting experiments on artificial data. We investigate the behavior of the following learning strategies:

view 1 Take a random consistent hypothesis (using a uniform distribution over the version space), ignoring all data from the second view.

view 2 Identical to the first strategy, but ignores the first view instead of the second.

rule R1, R2, R3 Proceed according to resolution rule R1, R2 or R3 from Section 5.4.1. The hypotheses h_1, h_2 are again chosen uniformly at random from the version space.

maj. vote If $(x_1, x_2) \in DIS_1 \times DIS_2$, vote for the label that holds the majority in the sample, i.e., make the exact opposite decision compared to resolution rule R3.

combined rule Proceed like the Combined Rule from Section 5.4.1.

In the experiment whose results are shown in Figure 5.8 we estimated the sample complexity under the worst-case distribution, as given in the proof of

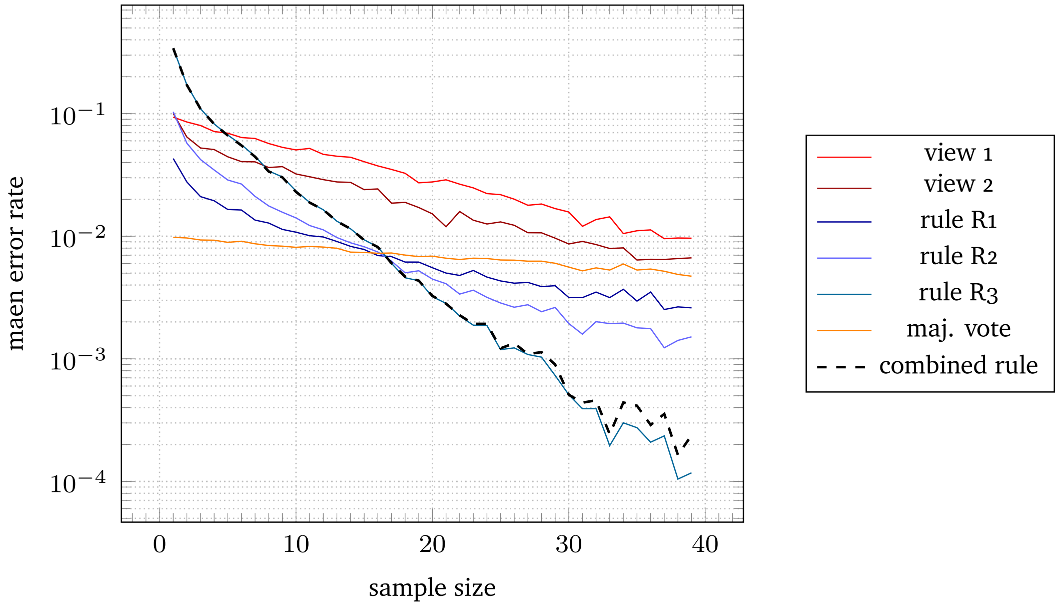


Figure 5.9: Estimated mean error rate under the worst-case distribution with the parameter $\varepsilon = 0.01$ for $(\mathcal{C}_1, \mathcal{C}_2) = (\mathrm{SF}_{10}, \mathrm{SF}_5)$.

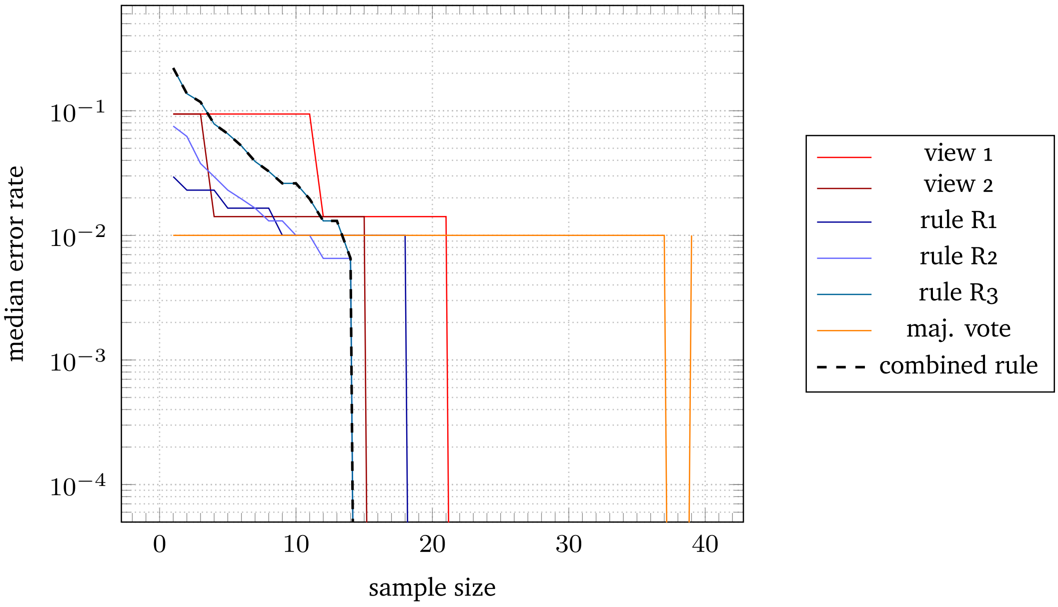


Figure 5.10: Estimated median error rate under the worst-case distribution with the parameter $\varepsilon = 0.01$ for $(\mathcal{C}_1, \mathcal{C}_2) = (\mathrm{SF}_{10}, \mathrm{SF}_5)$. The trimmed values below of 10^{-4} correspond to a median error rate of zero.

Lemma 55, for $p_+ = p_{min} = 1/(\theta_1\theta_2)$ and different values of ε on the classes $\mathcal{C}_1 = \mathrm{SF}_{10}$ and $\mathcal{C}_2 = \mathrm{SF}_5$.[8] For each learner and each value of ε, we increased the sample size (starting at $m = 1$) until at least one half of 500 independent runs of the learner, each time with a newly drawn target concept and sample, achieved an error rate smaller than ε. Figure 5.8 can verify several facts that we already derived from theory:

- The majority vote performs much worse than rule R3, which counterintuitively decides against the majority when in doubt. In fact, the experiment supports the conjecture that the majority vote is by far the worst of all learning strategies considered in this chapter when exposed to the worst-case distribution.
- Learning becomes harder with smaller values of ε. Even though all estimated sample complexities are much higher than the theoretical lower bound $m_0 = 1/40 \cdot \sqrt{n_1 n_2/\varepsilon}$ (with $n_1 = 8$ and $n_2 = 3$ in our experiments) from the proof of Lemma 55, the curves are indeed proportional to $\sqrt{1/\varepsilon}$.

We also observe that "view 2" performs better than "view 1", although the worst-case distribution balances the probabilities to uncover "interesting points" for both SF_{10} and SF_5 (see the proof of Lemma 55). A possible explanation is that the version space of SF_5 is typically smaller than the one of SF_{10} (simply because SF_5 contains fewer concepts while the fraction of consistent concepts declines similarly for both classes under the worst-case distribution) and the chance of guessing the target concept correctly is therefore higher for "view 2".

It can be surprising that the PAC-learners who ignore one view—and thereby forfeit the main advantage of the Co-training setting—perform quite well compared with the three resolution rules. We can see the advantage of Co-training by studying how the *mean error rates* fall with the increasing sample size in Figure 5.9. For this experiment, we used the worst-case distribution with the parameter ε set to $10^{-2} = 0.01$, which corresponds to the setting in Figure 5.8 on the very right, and estimated the mean error rate by 500 independent runs. As Figure 5.9 shows, the mean error rates of the resolution rules, especially R3 and the combined rule, fall much faster than the error rates of the one-view-PAC-learners.

The reader may be wondering how to reconcile Figures 5.8 and 5.9. For example, the former figure shows that "majority vote" requires about 40 labeled examples to achieve an error rate lower than 0.01 with a probability of at least one half, while the latter one demonstrates that "majority vote" has a mean error rate smaller than 0.01 for any sample size. The reason for the mismatch is

[8] We also run the experiment on $(\mathrm{SF}_{k_1}, \mathrm{SF}_{k_2})$ for different values of k_1 and k_2. The results are roughly equivalent, although the scale of the axes changes.

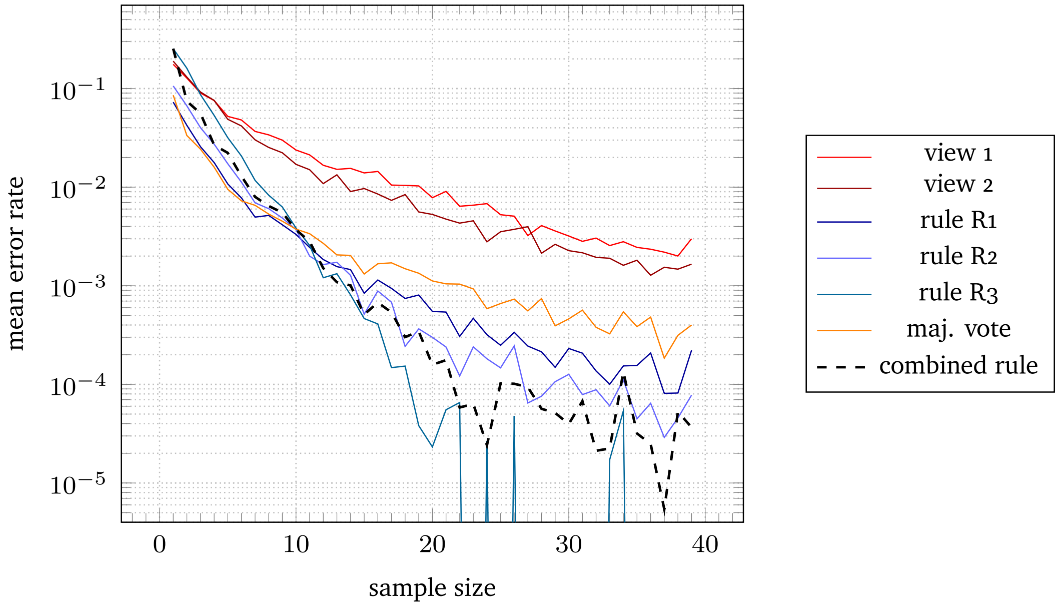


Figure 5.11: Estimated mean error rate in the "uniform" setting for $(\mathcal{C}_1, \mathcal{C}_2) = (\mathrm{SF}_{10}, \mathrm{SF}_5)$. The trimmed values below of 10^{-5} correspond to a mean error rate of zero.

the fact, that the distributions of the error rates are heavily skewed to different sides. We can verify this by plotting the *median error rates* (again estimated by 500 independent runs of the experiment), which are shown in Figure 5.10 and match nicely with the results on the very right of Figure 5.8, e.g., the median error rate of "majority vote" drops below 0.01 for the first time at a sample size of 38.

Furthermore, we can run experiments for other settings than the worst case. For example, Figure 5.11 depicts the mean error rates (estimated by 1000 independent runs) of the seven different learning strategies over the sample size in the following "uniform" setting: the target concept (c_1^*, c_2^*) is drawn uniformly at random from $(\mathrm{SF}_{10}, \mathrm{SF}_5)$, p^+ is drawn uniformly at random from $[0, 1]$ and $P(\cdot|+)$ and $P(\cdot|-)$ are uniform distributions. Since the parameter ε does not exist in this setting, we determined the input parameter ε for the Combined Rule empirically[9].

As expected, the error rates decrease faster than in the worst case (compare Figure 5.9). Note that the comparative behavior of the learning strategies is

[9] To this end we proceeded as follows. We split the sample into two equally sized parts: the first half for learning with different choices of ε, the second half as a test set to determine the empirically optimal ε. The resulting value was then used for learning on the whole sample.

similar, e.g., rule R3 and the Combined Rule again exhibit the fastest decreasing mean error rates.

5.6 Conclusions and Open Questions

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

The case of subclasses with infinitely many singletons The lower bound given in Theorem 59 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 59) 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 an open question.

Tight bounds for Co-training with k-tuples We saw in Section 5.4.5 that the generalized upper and lower bounds on the sample complexity leave a gap of $1/k$. It would be interesting to find bounds that are tight (up to logarithmic factors).

Since the learner in the k-tuple setting chooses k hypotheses h_b instead of just two, it is possible to design many new resolution rules (preferably with the constraint that our old three rules are chosen for the special case $k = 2$). We conjecture that the number of resolution rules necessary to prove tight bounds grows with k.

Agnostic learning Like in the previous chapters, the results presented in this chapter are only valid for the realizable case of PAC-learning. It is an open question if and how our results can be applied to more relaxed settings, like agnostic learning or learning with noise.

Chapter 6

An Example of Practical Application: Harnessing Unlabeled Data for an Attack on Audio CAPTCHAs

This chapter is based on the paper "Reducing the Cost of Breaking Audio CAPTCHAs by Active and Semi-supervised Learning" [DMK14][1], which is a joint work of the author with Hendrik Meutzner and Dorothea Kolossa. All results in this chapter were originally published in this paper[2].

6.1 Introduction

This chapter stands out from the rest of the thesis, in the sense that it is not concerned with proving theorems but solving a practical problem, namely performing an attack on a real-world internet security scheme. Nevertheless, we stay true to our main topic, which is the investigation of the power of unlabeled data, in particular the power of semi-supervised learning (we will also apply active learning in this chapter, but the main improvements, as we will see, originate from semi-supervised learning). We will show in this chapter, that unlabeled data can help considerably in real-world problems, while we have no good theory at hand that can explain or even predict these findings. It is likely that the good performance of semi-supervised learning on our problem

[1] This paper was supported in part by the DFG Research Training Group UbiCrypt (GRK 1817/1).

[2] A note from the author about the contributions of co-authors in this part of the thesis: most of the research was conducted jointly by Hendrik Meutzner and me. While Mr. Meutzner and Prof. Dr. Kolossa contributed the data sets and the ASR backend, I suggested the semi-supervised and active learning algorithms. Both Mr. Meutzner and I surveyed the literature, implemented the algorithms and run the experiments together. The final evaluation and interpretation of the experimental data was done primarily by me.

relied on extra assumptions, i.e., a connection between the data distribution P and the target concept (note that we consider “models” instead of “target concepts” in this chapter). Sadly we are not aware of any formulation of this assumption.

Let us now consider the problem we want to solve:

CAPTCHAs[3] are widely used in the Internet for distinguishing human users from automated programs to limit the abuse in online services, e.g., automated account creation for sending spam mail. Therefore, CAPTCHAs should be easy to solve by humans, but difficult to break by machines. Audio CAPTCHAs are required for allowing access to visually impaired users and to enable the use of non-graphical devices.

For evaluating the security strength of a CAPTCHA scheme, one can build a system that attempts to solve the CAPTCHAs in an automated manner. Recent investigations [TSHvA08, BBP$^+$11, SOO13] show that most audio CAPTCHAs are insecure as they can be broken by machine learning techniques. In this work, we utilize and improve an attack based on *automatic speech recognition* (ASR) pioneered by Sano, Otsuka and Okuno [SOO13].

Training speech recognition systems, as any machine learning technique, requires labeled data examples; thus to break audio CAPTCHAs by means of ASR, the orthographic transcription of the audio signals must be available. Whereas collecting unlabeled data (i.e., downloading audio CAPTCHAs from a website) is nearly for free, labeling the data is very expensive and time consuming as it requires human interaction.

Therefore we investigate the usefulness of unlabeled data, by means of active and semi-supervised learning, to reduce the amount of necessary labels. Our experiments are conducted on two different data sets, a development and an evaluation set. The development set is given by the Aurora-5 speech corpus [PH00] that contains thousands of noisy digit sequences. The digit sequences constitute real speech recordings spoken by various speakers. The evaluation set is given by the current version of reCAPTCHA [Goo14] that consists of highly distorted digit sequences where the speech appears to be synthetic.

6.1.1 Related Work

The two-stage approach from breaking visual CAPTCHAs

While there is much prior work on breaking visual CAPTCHAs (e.g., see [CS04, CLSC05, HGH08, YEA07]), less focus has been given to audio CAPTCHAs. Chellapilla, et al. establish in [CLSC05] that most attacks on character-based visual CAPTCHAs employ a two-stage approach: the image is first segmented

[3] Completely Automated Public Turing tests to tell Computers and Humans Apart

into the parts where characters are located and then the individual characters are recognized using standard pattern recognition techniques. Chellapilla et al. consider the segmentation phase to be the hard part of the problem. A two-stage approach is also applied successfully to break audio CAPTCHAs by Tam et al. [TSHvA08] and Bursztein et al. [BBP$^+$11].

Automatic speech recognition (ASR)

The viability of ASR to break audio CAPTCHAs was recently demonstrated by Sano, Otsuka and Okuno in [SOO13]. Since ASR can cope with continuous speech, the aforementioned segmentation problem is avoided. For an introduction to ASR we refer the interested reader to Rabiner and Juang [RJ93].

Semi-supervisead learning for ASR

Our semi-supervised learning algorithm is an instance of the *hard expectation-maximization (EM) algorithm* [DLR77, KMN97], which was previously applied for ASR by Zavaliagkos and Colthurst in [ZC98], for the similar problem of sequence classification using Hidden Markov Models by Zhong in [Zho05] and in the related field of computational linguistics by Spitkovsky et al. in [SAJM10].

Some remarks about the hard EM algorithm are in order: the well-known *expectation-maximization (EM) algorithm*, introduced by Dempster, Laird and Rubin in [DLR77], iteratively finds a (local) maximum likelihood estimation for models depending on hidden variables, like the unknown labels of unlabeled CAPTCHAs in our case. In each iteration the algorithm maximizes the expected value of the likelihood function with respect to the distribution over the hidden variables conditioned on the model from the last iteration. The *hard* EM algorithm, a term coined by Kearns, Mansour and Ng in [KMN97], is a simplification of standard EM, where the distribution is replaced by the mode, i.e., the most likely value, of the hidden variables. Hard EM is sometimes called *Viterbi EM* (e.g., see [SAJM10]); in this chapter we will always refer to the Viterbi algorithm for approximating the most likely path in a Hidden Markov Model [Vit67], which is also an instance of the hard EM algorithm and heavily used in ASR, when we write the name “Viterbi”.

Combining active and semi-supervised learning for ASR

The active learning algorithm that we employ as well as a method of combining active and semi-supervised learning in parallel were originally proposed in the context of ASR by Riccardi and Hakkani-Tür [RHT03]. Similar combinations of active and semi-supervised learning for speech recognition were also investigated by Yu et al. [YVDA10].

Our serial combination of active and semi-supervised learning is—to our knowledge—new in the field of ASR.

6.1.2 Main Results

We demonstrate that the labeling costs can be reduced considerably by both semi-supervised and active learning while achieving high success rates for breaking audio CAPTCHAs. We will see that especially semi-supervised learning provides a huge saving of labeled data.

The investigation of combinations of semi-supervised and active learning will show that the parallel method from the ASR literature [RHT03] performs much worse than a considerably simpler (conceptually and computationally) serial approach, which—to our knowledge—was not considered before in the field of ASR. We will show how the parallel method can still yield good results by introducing weights, although adjusting these weights seems to be too expensive (when the cost is measures in labeled data) for practical attacks.

We show that results on the development set predominantly translate to the evaluation set, while analyzing differences in performance between the two data sets can give insights into the design of CAPTCHAs that are harder to break by active learning.

Although we put our focus on the comparison of different methods, and not on achieving the smallest possible error rate on labeling unseen CAPTCHAs, the we in fact attain a smaller error rate (of 39%) than the best error rate (of 83%, see [SOO13]) in the published literature (with training performed on a similar amount of data).

6.1.3 Organization of This Chapter

First, we will define some vocabulary in Section 6.2, which is used in the remainder of the chapter. In Section 6.3, we introduce our four learning algorithms: semi-supervised, active, and two combined learners. The setup and results of our experiments are given in Section 6.4. In Section 6.5 with some suggestions for the design of audio CAPTCHAs, that are stronger against active learning attacks. The chapter is finished with conclusions and open questions in Section 6.6.

6.2 Preliminaries

An *observation* is a representation of an audio signal (by an appropriate sequence of feature vectors), that consists of a variable number of spoken words, which are possibly separated by silence or background noise. The vocabulary is limited to digits between zero and nine and the speech may be either natural or synthetic. The observations may contain additional distortions, e.g., additive noise or reverberation effects. The *label* of an observation is the transcription (including silence and noise) of the audio signal. Note that, unlike in the previous chapters, labels are no longer binary, instead they are strings over the alphabet $\Sigma = \{0, \ldots, 9, \text{noise}\}$ (where “noise” also denotes silence).

A *model* represents a data structure that enables the estimation of labels—together with a measure of confidence for each estimated label—for a given set of observations. We regard an estimated label to be *correct* if its digit sequence is identical to the ground-truth label of the observation disregarding any occurrence of silence or noise.[4] Models used in ASR are often generative, i.e., they describe a distribution over observation-label pairs. The probability of a (fixed) model generating an observation-label pair is called the *likelihood* of that pair.

In this chapter, a *learner* is an algorithm that takes a sample of labeled observations as its input and outputs a model. The sample is drawn according to some unknown distribution (which is not necessarily generated by a model). We measure the error rate of a learner's model according to the same distribution.

When comparing two learners L and L', we call L *better* than L' if L achieves a lower average error rate for any fixed number of labeled words in the sample. Another possible criterion is based the number of labeled observations instead of words, which is unsuited for our purpose as observations can vary in length and, clearly, the effort of labeling an observation is proportional to its word count.

We assume that we have access to an ASR system that can learn models using labeled observations and that can estimate labels together with a measure of confidence for unlabeled observations. Please note that our learning algorithms, which are defined in the next section, are oblivious to the implementation of this underlying ASR backend (for a description of the backend used in our experiments see Sections 6.3.4 and 6.3.5).

6.3 Algorithms

6.3.1 Semi-supervised Learning

In semi-supervised learning the learner has access to two samples: a labeled sample (which we denote by S_L) and an unlabeled one (denoted by S_U). Both

[4] For breaking reCAPTCHA this measure could be even more generous, as the reCAPTCHA system accepts answers having a Levenshtein distance of one between the user input and the true label of the CAPTCHA.

samples are drawn according to the same distribution over the observations, thus we can regard the sample S_U as providing additional information about the distribution to the learner. A detailed introduction to semi-supervised learning is given in [CSZ06].

We propose the semi-supervised learning algorithm from Listing 6.1, which—as mentioned before—has already been applied in similar settings [ZC98, Zho05, SAJM10]. In this, as in all following iterative learning algorithms, the number of iterations can be either determined by convergence or chosen in advance by the user.

Algorithm 6.1 Semi-supervised learner

```
1. Train initial model $M$ on a (small) randomly drawn labeled sample $S_L$.
2. Draw an unlabeled sample $S_U$.
3. Repeat the following steps:
    a) Estimate the labels of $S_U$ using the current model $M$.
    b) Retrain model $M$ on the labeled set $S_L \cup S_U$.
```

In [ZC98] Zavaliagkos and Colthurst removed the estimated labels with the least confidence scores from S_U before retraining the model; we will use a similar idea for the learning algorithms in section 6.3.3, but instead of merely removing the estimations we will apply active learning to obtain the true labels.

6.3.2 Active Learning

An active learner, introduced by Cohn el al. in [CAL94], also has access to an unlabeled sample, but here the learner is allowed to actively request true labels for selected observations.

The active learner from Listing 6.2 requests labels for those observations which the learner's model is least confident about.

The difference between our method and the active learner from Riccardi and Hakkani-Tür [RHT03] is that the latter one uses one pool of unlabeled observations, while we draw a new sample S_U in every iteration. Moreover, we use a different measure of confidence in our experiments than Riccardi and Hakkani-Tür (see Section 6.3.5).

As noticed in [RHT03], the choice of n is a trade-off between making optimal selection decisions (n should be small) and computational cost (n should be large).

Algorithm 6.2 Active learner

```
1. Train initial model $M$ on a (small) randomly drawn labeled sample $S_L$.
2. Repeat the following steps:
    a) Draw an unlabeled sample $S_U$ and estimate its labels using the
       current model $M$.
    b) Request the true labels of the $n$ observations from $S_U$ with the least
       confidence and add them to $S_L$.
    c) Retrain model $M$ on the labeled set $S_L$.
```

6.3.3 Combined Learners

Since active learning algorithms typically do not request labels for all drawn unlabeled observations, it is a natural idea to use these leftovers for semi-supervised learning.

We study two methods to combine the active learner from Section 6.3.2 with the semi-supervised learner from Section 6.3.1.

Our first combined learner is given in Listing 6.3. This algorithm was proposed by Riccardi and Hakkani-Tür in [RHT03]—apart from the parts marked as 'optional'—and operates in *parallel*, i.e., in each iteration of the active learner we use the spare unlabeled data for semi-supervised learning.

Algorithm 6.3 Combined learner, parallel

```
1. Train initial model $M$ on a (small) randomly drawn labeled sample $S_L$.
2. Set $S_U$ to the empty set.
3. Repeat the following steps:
    a) Draw an unlabeled sample $S'_U$ and estimate its labels using the
       current model $M$.
    b) Request the true labels of the $n$ observations of $S'_U$ with the least
       confidence and add them to $S_L$. Add the remaining observations of
       $S'_U$ with their estimated labels to $S_U$.
    c) Retrain model $M$ on the labeled set $S_L \cup S_U$ (optional: give $S_L$ a
       higher weight than $S_U$)[5].
    d) Optional: Iterate the semi-supervised learning steps, i.e., estimate
       the labels of $S_U$ and retrain $M$ on $S_L \cup S_U$ several times.
```

During our experiments we had to include weighting and additional semi-supervised steps to achieve satisfactory results. The benefit of additional semi-supervised learning steps was previously recognized by Yu et al. [YVDA10].

For our second combined method, given in Listing 6.4, we propose to run the active and semi-supervised learners in a *serial* fashion.

Algorithm 6.4 Combined learner, serial

1. Train model M using the active learner from Listing 6.2
2. Use M as the initial model for the semi-supervised learner from Listing 6.1 (where the first step is omitted).

In spite of the simplicity of this "folklore" algorithm, we are not aware of prior applications in the field of automatic speech recognition.

The serial combined learner is computationally simpler than the parallel one, especially if the optional additional semi-supervised learning steps are executed. For a discussion of the parallel learner's practicability see Section 6.4.2.

6.3.4 Speech Recognition Backend

For the underlying ASR backend we employ the *Hidden Markov Model Toolkit* (HTK) [YKO$^+$06], which provides a state-of-the-art implementation of the required training and recognition algorithms used for our approach.

We use the Baum-Welch algorithm [BPSW70] to train models given a labeled set of observations, and the estimation of labels—given the trained models—is based on the Viterbi algorithm [Vit67]. Due to space constrains we omit the details of these algorithms and refer the interested reader to Rabiner and Juang [RJ93]. HTK's implementation of the Viterbi algorithm, based on a token passing model, provides a list of the n most likely labels together their corresponding word likelihoods, and the overall likelihood of each label can be simply computed as the product of the individual word likelihoods.

6.3.5 Measures of Confidence

Let s be an unlabeled observation and let $(l_1, \boldsymbol{p}_1), \ldots, (l_n, \boldsymbol{p}_n)$ be the output of the Viterbi algorithm using model M, i.e., l_i are the labels estimated by M for s and $\boldsymbol{p}_i$ are the associated word likelihood vectors (including likelihoods for parts consisting of silence or noise). We assume that the labels are ordered by their likelihoods.[6] We define the *normalized likelihood* q_i of $(l_i, \boldsymbol{p}_i)$ as

$$q_i := \left(\prod_{k=1}^{|l_i|} p_{i,k}^{\frac{1}{|w_{i,k}|}} \right)^{\frac{1}{|l_i|}}$$

where $p_{i,k}$ is the likelihood of the k-th word of label l_i, $|w_{i,k}|$ is its length (in frames) and $|l_i|$ is the word count of l_i.

While we always choose l_1 to be the estimated label of s for semi-supervised learning, we employ these two methods to measure the confidence in l_1 for active learning:

(A) Take the *normalized likelihood* q_1 of $(l_1, \boldsymbol{p}_1)$.

(B) Let i be the first index such that l_1 and l_i differ in more than just the positions of silence/noise. Compute the confidence as the ratio $q_1 \,/\, q_i$. We call this quantity the *likelihood ratio*.

We give the following intuition why the normalized likelihood is a better measure of confidence than the plain likelihood: normalizing by $1/|w_{1,k}|$ counteracts against choosing observations with predominantly long words. The normalization by $1/|l_1|$ prevents the learner from selecting observations with a high word count, which would be a good strategy, in fact, if the effort of labeling was not directly dependent on the word count.

Using the likelihood ratio instead of the normalized likelihood is helpful in the following situation: it can happen that the model M is still unsure about the position of silence/noise in s, while the digits are already very certain; let us say the digits are in the same order in $l_1, \ldots, l_m$ for some m. Let s' be a second observation where M has a low level of certainty about the order of digits: let $(l'_1, \boldsymbol{p_1}'), \ldots, (l'_n, \boldsymbol{p_n}')$ be the output of the Viterbi algorithm for s' and let us say that already l'_1 and l'_2 differ in their digits and $q'_1 \approx q'_2$. It can occur in this case that q_1 is smaller than q'_1, so the active learner based on the normalized likelihood would request the label of s first. But we would like the learner to request the true label of s' instead of s, because we are only interested in the order of digits when breaking CAPTCHAs. Taking the quotient can mitigate this problem, since typically q_{m+1} is much smaller than q_1, or it may even hold that $m = n$, in which case we consider the likelihood quotient to be infinite.

For example, during our experiments the Viterbi algorithm estimated the following labels for an observation with the true label “noise-7-2-6-9-noise”:

- “noise-7-2-6-9-noise”
- “noise-noise-7-2-6-9-noise”

[6] i.e., it holds $\prod_{k=1}^{|l_i|} p_{i,k} \geq \prod_{k=1}^{|l_j|} p_{j,k}$ for $i < j$.

- “noise-7-2-6-9-noise-noise”
- “noise-noise-7-2-6-9-noise-noise”
- “noise-7-2-6-noise-9-noise’
- etc.

We computed the sixteen most likely labels and all of them were correct, i.e. only differing in the position and number of noise segments. However, the normalized likelihood q_1 of this observation was lower than the normalized likelihood of other, still erroneously labeled observations.

6.4 Experiments

6.4.1 Data Sets

We use two different data sets: Aurora-5 [PH00] and a sample from Google’s current audio CAPTCHA scheme (reCAPTCHA) [Goo14].

The Aurora-5 data set contains real speech recordings that are mixed with natural background noise (e.g., airport, car engine or office noise) at different *signal-to-noise ratios* (SNRs). The speech recordings were obtained from different speakers (approx. 50 males and 50 females), each pronouncing several observations that comprise a sequence of digits and the number of digits per observation is varied between one and seven. The data set is provided with the respective labels for each observation. Our experiments are based on the noisy digit recordings distorted by the *interior noise* at 10 dB SNR.

We find that using Aurora-5 for system development is advantageous as the recordings are not only similar to digit-based audio CAPTCHAs but they offer the advantage of having access to a large pool of 9275 correctly labeled observations with 50554 words.

The reCAPTCHA data set is also constructed from a sequence of digits where the total number of digits per observation is varied between 6 and 12. The digits are spoken by both a male and a female voice and the speech appears synthetic. All signals exhibit the same stationary background noise. The CAPTCHAs were downloaded from the reCAPTCHA website in March 2014 and manually labeled at our institute. To obtain the CAPTCHA labels, we conducted a listening test in a controlled environment involving a group of 12 human listeners. Each participant was asked to label a set of 50 CAPTCHA signals. The participants were briefed that the signals consist of a varying number of digits that are separated in time by distinct speech pauses. Each CAPTCHA was labeled by four different participants to identify inconsistent labelings. We only utilize those transcriptions that exhibit an agreement between at

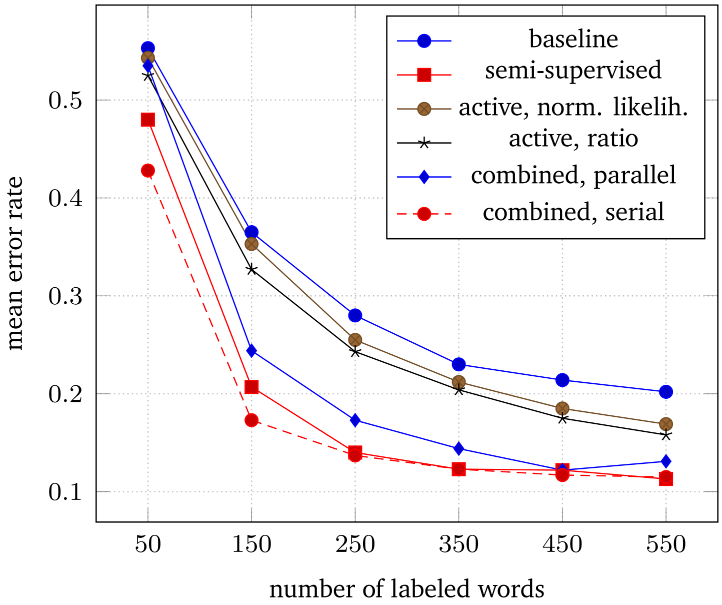


Figure 6.1: Mean error rates for Aurora-5.

	baseline		**semi-supervised learner**	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0– 99	0.553	0.103	0.480	0.187
100–199	0.365	0.071	0.207	0.094
200–299	0.280	0.051	0.140	0.036
300–399	0.230	0.041	0.123	0.032
400–499	0.214	0.036	0.122	0.026
500–599	0.202	0.035	0.113	0.020

	active, normalized likelihood		**active, likelihood ratio**	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0– 99	0.543	0.109	0.525	0.119
100–199	0.353	0.065	0.327	0.067
200–299	0.255	0.039	0.243	0.035
300–399	0.212	0.030	0.204	0.036
400–499	0.185	0.025	0.175	0.025
500–599	0.169	0.026	0.158	0.019

	parallel combined learner		**serial combined learner**	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0– 99	0.535	0.173	0.428	0.170
100–199	0.244	0.120	0.173	0.049
200–299	0.173	0.042	0.137	0.031
300–399	0.144	0.034	0.123	0.027
400–499	0.122	0.023	0.117	0.027
500–599	0.131	0.030	0.115	0.032

Table 6.1: Experimental results for Aurora-5.

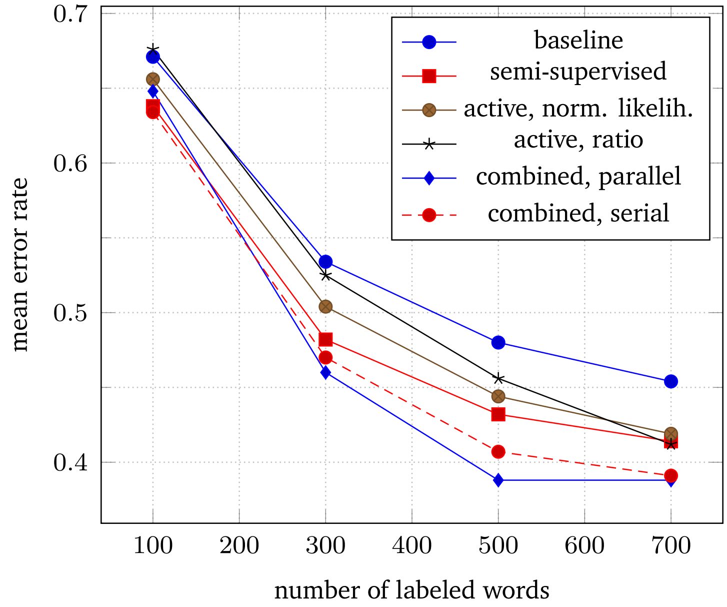


Figure 6.2: Mean error rates for reCAPTCHA 2014.

	baseline		semi-supervised learner	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0–199	0.671	0.104	0.638	0.144
200–399	0.534	0.098	0.482	0.136
400–599	0.480	0.100	0.432	0.136
600–799	0.454	0.103	0.414	0.135

	active, normalized likelihood		active, likelihood ratio	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0–199	0.656	0.114	0.676	0.112
200–399	0.504	0.093	0.525	0.098
400–599	0.444	0.094	0.456	0.098
600–799	0.419	0.088	0.412	0.093

	parallel combined learner		serial combined learner	
# words	mean error rate	std. dev.	mean error rate	std. dev.
0–199	0.648	0.125	0.634	0.147
200–399	0.460	0.111	0.470	0.116
400–599	0.388	0.105	0.407	0.124
600–799	0.388	0.108	0.391	0.118

Table 6.2: Experimental results for reCAPTCHA.

least three participants. This procedure resulted in an overall number of 336 labeled CAPTCHAs—corresponding to 4859 labeled words—that we utilize for our evaluation.

An important difference between the two data sets arises from the speaking pauses that separate the individual digits from each other. While for the Aurora-5 observations there are speaking pauses between all digits, the reCAPTCHA observations consists of blocks of digits that are spoken in an unnaturally fast succession. We assume that this is done to make the separation problem harder.

6.4.2 Setup and Learning

To examine and compare the performance of our learning algorithms we conducted a series of experiments. Because we put the focus on comparison, we did not try to thoroughly optimize parameters shared by all learners to achieve the best possible error rate.

However, we were able to reduce the error rate for solving the current reCAPTCHA challenges exactly (i.e., with a Levenshtein distance of zero) by 44%—from 83%, published by Sano, Otsuka and Okuno [SOO13], to 39%—with a similar amount of training data.[7]

In this paragraph we specify details of our ASR backend (these details are not required for understanding the rest of the chapter; again, we refer interested readers unfamiliar with ASR to Rabiner and Juang [RJ93]). We used 39-dimensional *Mel frequency cepstral coefficients* (MFCCs) as features including their first and second order derivatives, where each feature vector corresponds to a window length of 25 ms of the audio signal. Each digit is modeled by an HMM that has 18 states and exhibits a left-to-right topology without state skips. The silence/noise model had five states and allowed backward transitions and skips between the first and the last state. The state emission probabilities were represented by a Gaussian mixture model (GMM) having four mixture components.

In all experiments, the ratio between the number of randomly drawn observations (labeled and unlabeled) and the number of labels used was kept constant. On the Aurora-5 data set this ratio is set to ten and on reCAPTCHA, due to the small amount of available labeled data, it is four. Intuitively, for optimal error rates this ratio should be as large as possible, but for comparisons any constant suffices. All iterations were done five times. Initial models were trained on fifty (Aurora-5) resp. twenty (reCAPTCHA) randomly chosen labeled observations.

[7] The error rates in [SOO13] were found to be 83% and 48%, for a Levenshtein distance of zero and one, respectively, with 400 labeled observations available in total.

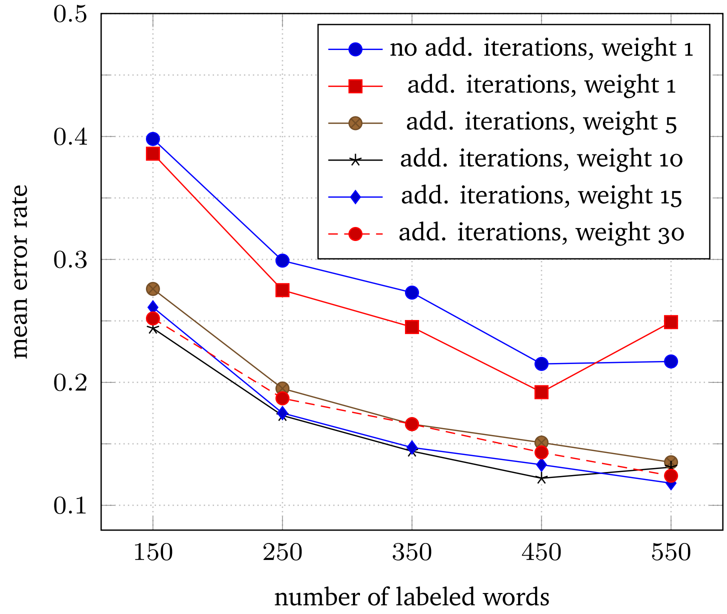


Figure 6.3: Mean error rates for variants of the parallel combined learner using different weighting factors and options of additional semi-supervised learning iterations on Aurora-5.

Baseline

In addition to the learning algorithms given in Section 6.3, we also run a purely supervised learner as a baseline. The baseline learner started—as all our learning algorithms—with an initial model and then added an increasing number of randomly drawn labeled observations using Baum-Welch training. We kept track of the number of additional labeled words and the empirical error rate (evaluated on an independent test set). After 10 (Aurora-5) resp. 20 (reCAPTCHA) runs, we calculated the mean error rates and standard deviations for several intervals of word counts. In the other experiments we proceeded in the same way and the results are shown in Table 6.1 (Aurora-5) resp. Table 6.2 (reCAPTCHA). A plot of the mean error rate against the word count is given in Figure 6.1 resp. Figure 6.2.

Semi-supervised learner

The semi-supervised learner operates in a similar fashion: it uses the model trained by the baseline as its initial model and then applies five rounds of semi-supervised training on randomly drawn unlabeled data (with a constant ratio between the number of unlabeled and labeled data as outlined above). As shown in Table 6.1 resp. Table 6.2, the semi-supervised learner clearly outperforms the baseline, although the standard deviation is higher when the number of additional words is small.

A comparison with the supervised baseline allows us to estimate the value of unlabeled data: for example, on the Aurora-5 data set, we get a mean error rate of 21% by improving the initial model with 150 labeled and ca. 1350 unlabeled words[8], however we need over 500 labeled words to achieve the same performance when no unlabeled data are available.

Active learners

We examined both measures of confidence from Section 6.3.5 for active learning. As you can see in Figure 6.1 and 6.2, both active learners are better than the baseline, which implies that both confidence measures can indeed identify observations whose label is more informative than the label of a randomly chosen one. Here we notice an important difference in the two data sets: on Aurora-5 the confidence measure based on the likelihood ratio outperforms the measure based on the normalized likelihood, while on reCAPTCHA we see the opposite behavior (except for a word count above 600, where the quotient is slightly better, which we interpret as an outlier.). The implications of this observation are discussed in Section 6.5.

Combined learners

The combined learners are based on the superior confidence measure for each data set. We find that both parallel and serial learners are better than pure active learning, and on each data set one of the two combined learners is the best learner evaluated in this work. Again, the use of semi-supervised learning increases the standard deviation for small word counts.

Practicability and performance of the parallel combined learner

As seen in Figure 6.3, we had to introduce weighting (with an empirically found optimal weight of ten for Aurora-5 and six for reCAPTCHA) and additional semi-supervised iterations to achieve a satisfactory performance for the parallel learner from Riccardi and Hakkani-Tür [RHT03]. Please note that optimizing a parameter by experimentation—and by similar techniques like cross validation—is detrimental to active learning's goal of reducing the number of used labels, because an active learner will (typically) request the labels of a different fraction of the unlabeled data pool in each run. We are not aware

[8] Since on the Aurora-5 data set the ratio between all used observations and labeled observations is ten, and we designed the active learner in such a way that it has no bias in observation length.

of a method for optimizing the weight without using a large amount of labeled data, which is not readily available in situations where active learning is applied.

Furthermore, the optimized parallel learner is still worse than pure semi-supervised learning on Aurora-5, which implies that its active choice of labels is inferior to random sampling. While its performance on reCAPTCHA is better than the serial learner (except for small word counts), this gain seems too small to us to be worth the effort in real world attacks.

6.5 Implications for the Design of CAPTCHAs

The observation that active learning based on the likelihood ratio performs better than the one based on the normalized likelihood on the Aurora-5 data set, is best explained by the frequent confusion about the position of noise or silence (the example at the end of Section 6.3.5 was from Aurora-5). Further evidence is given by the fact that the normalized likelihood performs better on the reCAPTCHA data set, where we did not see these kind of confusions.

The computation of likelihood ratios is more expensive, since we have to determine the n best hypotheses instead of just one. To improve the resistance of audio CAPTCHAs against active learning, we would like to see this gap between the performances of the two methods on audio CAPTCHAs, and, at best, the gap should be much larger.

We conjecture that both can be achieved by inserting superfluous words[9] from a preferably large vocabulary between the words that are relevant for solving the CAPTCHA (e.g., the digits in reCAPTCHA). These superfluous words play the role of the noise in Aurora-5.

Doing so should amplify the effect that occurred on Aurora-5: one simple silence/noise model does not suffice anymore; in the best case a different model must be trained for each superfluous word (the existence of a small number of models that recognize all superfluous words well must be prevented by the designer of the CAPTCHA scheme). Since the Viterbi algorithm must estimate which, where and how many superfluous words occurred, the confusion in the n best hypotheses will be much higher than it is now for Aurora-5. Therefore, even if the digit sequence is already very certain, the estimated labels will be assigned a low likelihood, and it becomes harder to distinguish already correctly recognized from uncertain digit sequences.

The introduction of superfluous words should not hamper human users too much, because they can focus on the necessary words (e.g., digits). On the

[9] These do not necessarily have to be English words. Even noise signals could be used for the purpose.

other hand, the separation problem still has to be hard, so transitions between necessary and superfluous words have to be hard to detect.

6.6 Conclusions and Open Questions

We close this chapter with the following open questions:

- There are many alternative ways to measure the confidence in an estimated label than the two options that we considered. It could be worthwhile to experiment with more measures. Furthermore, as we saw a difference between Aurora-5 and reCAPTCHA, it would interesting to gain more insight how to choose a good measure depending on the data set (without using much labeled data).

- We ruled out the parallel combination of the semi-supervised and active learner for practical attacks because we need (too) many labels to adjust the necessary weight. It would be interesting to prove sample size lower bounds for parameter optimizations like these. Alternatively, one could try to find efficient optimization algorithms, that do not require many labels.

- We conjectured in Section 6.5 how one could make attacks based on active learning techniques (like the ones that we applied) harder. Try to prove or disprove this conjecture (e.g., by designing a corresponding CAPTCHA scheme and evaluating our methods)!

- The question how well our results transfer to other audio CAPTCHA schemes is open.

Bibliography

[Ale86] Kenneth S. Alexander. Sample Moduli for Set-indexed Gaussian Processes. *The Annals of Probability*, 14(2):598–611, 04 1986.

[Ale87] Kenneth S. Alexander. Rates of Growth and Sample Moduli for Weighted Empirical Processes Indexed by Sets. *Probability Theory and Related Fields*, 75(3):379–423, 1987.

[AO07] Peter Auer and Ronald Ortner. A New PAC Bound for Intersection-closed Concept Classes. *Machine Learning*, 66:151–163, March 2007.

[BB10] Maria-Florina Balcan and Avrim Blum. A Discriminative Model for Semi-supervised Learning. *Journal of the ACM*, 57(3):19:1–19:46, 2010.

[BBL09] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic Active Learning. *Journal of Computer and System Sciences*, 75(1):78–89, January 2009.

[BBP$^+$11] Elie Bursztein, Romain Beauxis, Hristo S. Paskov, Daniele Perito, Celine Fabry, and John C. Mitchell. The Failure of Noise-based Non-continuous Audio Captchas. In *IEEE Symposium on Security and Privacy*, pages 19–31, 2011.

[BBY04] 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, NIPS*, 2004.

[BBZ07] Maria-Florina Balcan, Andrei Z. Broder, and Tong Zhang. Margin Based Active Learning. In *Proceedings of the 20th Annual Conference on Learning Theory, COLT*, pages 35–50, 2007.

[BDLP08] 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 21th Annual Conference on Learning Theory, COLT*, 2008.

[BEHW87] Alselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's Razor. *Information Processing Letters*, 24(6):377–380, April 1987.

[BEHW89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BHV10] Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The True Sample Complexity of Active Learning. *Machine Learning*, 80(2-3):111–139, 2010.

[BI91] Gyora M. Benedek and Alon Itai. Learnability with Respect to Fixed Distributions. *Theoretical Computer Science*, 86, 1991.

[BL13] Maria-Florina Balcan and Philip M. Long. Active and Passive Learning of Linear Separators under Log-concave Distributions. In *Proceedings of the 26th Annual Conference on Learning Theory, COLT*, pages 288–316, 2013.

[BM98] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the 11th Annual Conference on Learning Theory, COLT*, pages 92–100, 1998.

[BPSW70] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 02 1970.

[CAL94] David Cohn, Les Atlas, and Richard Ladner. Improving Generalization with Active Learning. *Machine Learning*, 15(2):201–221, 1994.

[Che52] Herman Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 12 1952.

[Chu74] Kai Lai Chung. *A Course in Probability Theory*. Probability and Mathematical Statistics. Academic Press, 1974.

[CLSC05] Kumar Chellapilla, Kevin Larson, Patrice Y. Simard, and Mary Czerwinski. Building Segmentation Based Human-Friendly Human Interaction Proofs (HIPs). In *Human Interactive Proofs*, Lecture Notes in Computer Science, pages 1–26, 2005.

[CS04] Kumar Chellapilla and Patrice Y. Simard. Using Machine Learning to Break Visual Human Interaction Proofs (HIPs). In *Advances in Neural Information Processing Systems 17, NIPS*, 2004.

[CSZ06] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-supervised Learning*. MIT Press, 2006.

[Dar15] Malte Darnstädt. The Optimal PAC Bound for Intersection-closed Concept Classes. *Information Processing Letters*, 115(4):458 – 461, 2015.

[Das05] Sanjoy Dasgupta. Coarse Sample Complexity Bounds for Active Learning. In *Advances in Neural Information Processing Systems 18, NIPS*, 2005.

[DHM07] Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A General Agnostic Active Learning Algorithm. In *Advances in Neural Information Processing Systems 20, NIPS*, 2007.

[DKRZ94] Richard M. Dudley, Sanjeev R. Kulkarni, Thomas J. Richardson, and Ofer Zeitouni. A Metric Entropy Bound is not Sufficient for Learnability. *IEEE Transactions on Information Theory*, 40(3):883–885, 1994.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[DMK14] Malte Darnstädt, Hendrik Meutzner, and Dorothea Kolossa. Reducing the Cost of Breaking Audio CAPTCHAs by Active and Semi-supervised Learning. To appear in the Proceedings of the 13th International Conference on Machine Learning and Applications, ICMLA, 2014.

[DS11] Malte Darnstädt and Hans Ulrich Simon. Smart PAC-Learners. *Theoretical Computer Science*, 412(19):1756 – 1766, 2011.

[DSS13] Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Unlabeled Data Does Provably Help. In *30th International Symposium on Theoretical Aspects of Computer Science, STACS*, volume 20, pages 185–196, 2013.

[DSS14] Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Supervised Learning and Co-training. *Theoretical Computer Science*, 519:68–87, January 2014.

[EHKV89] Andrzej Ehrenfeucht, David Haussler, Michael J. Kearns, and Leslie G. Valiant. A General Lower Bound on the Number of Examples Needed for Learning. *Information and Computation*, 82(3):247–261, 1989.

[Fel68] William Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, New York, 1968.

[GG89] Mihály Geréb-Graus. *Lower Bounds on Parallel, Distributed and Automata Computations*. PhD thesis, Harvard University Cambridge, MA, USA, 1989.

[GK06] Evarist Giné and Vladimir Koltchinskii. Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[Goo14] Google. reCAPTCHA, March 2014. http://www.recaptcha.net as of 03/2014.

[Han07] Steve Hanneke. A Bound on the Label Complexity of Agnostic Active Learning. In *Proceedings of the 24th international conference on Machine learning (ICML)*, pages 353–360, 2007.

[Han09] Steve Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

[HGH08] Abram Hindle, Michael W. Godfrey, and Richard C. Holt. Reverse Engineering CAPTCHAs. In *15th Working Conference on Reverse Engineering, WCRE*, pages 59–68, Oct 2008.

[Kää05] Matti Kääriäinen. Generalization Error Bounds Using Unlabeled Data. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT*, 2005.

[KG71] Julian Keilson and Hans-Ulrich Gerber. Some Results for Discrete Unimodality. *Journal of the American Statistical Association*, 66:386–389, 1971.

[KMN97] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. An Information-theoretic Analysis of Hard and Soft Assignment Methods for Clustering. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, UAI*, pages 282–293, 1997.

[KV94] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA, 1994.

[Mor94] Peter Morris. *Introduction to Game Theory*. Universitext. Springer New York, 1994.

[PH00] David Pearce and Hans-Günter Hirsch. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In *Proceedings of Interspeech*, pages 29–32, 2000.

[RHT03] Giuseppe Riccardi and Dilek Z. Hakkani-Tür. Active and Unsupervised Learning for Automatic Speech Recognition. In *Proceedings of Interspeech*, 2003.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[SAJM10] Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky, and Christopher D. Manning. Viterbi Training Improves Unsupervised Dependency Parsing. In *Proceedings of the 14th Conference on Computational Natural Language Learning, CoNLL*, pages 9–17, 2010.

[Sau72] Norbert Sauer. On the Density of Families of Sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[Sim09] Hans Ulrich Simon. Smart PAC-Learners. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory, ALT*, pages 353–367, 2009.

[SNZ08] Aarti Singh, Robert D. Nowak, and Xiaojin Zhu. Unlabeled Data: Now it Helps, Now it Doesn't. In *Advances in Neural Information Processing Systems 21, NIPS*, 2008.

[SOO13] Shotaro Sano, Takuma Otsuka, and Hiroshi G. Okuno. Solving Google's Continuous Audio CAPTCHA with HMM-based Automatic Speech Recognition. In *International Workshop on Security, IWSEC*, pages 36–52, 2013.

[TSHvA08] Jennifer Tam, Jirí Simsa, Sean Hyde, and Luis von Ahn. Breaking Audio CAPTCHAs. In *Advances in Neural Information Processing Systems 21, NIPS*, pages 1625–1632, 2008.

[Val84] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[VC71] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

[Vit67] Andrew J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[vN28] John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

[War04] Manfred K. Warmuth. The Optimal PAC Algorithm. In *Proceedings of the 17th Annual Conference on Learning Theory, COLT*, pages 641–642, 2004.

[WZ10] Wei Wang and Zhi-Hua Zhou. A New Analysis of Co-training. In *Proceedings of the 27th International Conference on Machine Learning, ICML*, pages 1135–1142, 2010.

[Yao77] Andrew Chi-Chin Yao. Probabilistic Computations: Toward a Unified Measure of Complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science, SFCS*, pages 222–227, Washington, DC, USA, 1977.

[YEA07] Jeff Yan and Ahmad S. El Ahmad. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms. In *Proceedings of the 23rd Annual Computer Security Applications Conference, ACSAC*, pages 279–291, Dec 2007.

[YKO$^+$06] Steve J. Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[YVDA10] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero. Active Learning and Semi-supervised Learning for Speech Recognition: A Unified Framework Using the Global Entropy Reduction Maximization Criterion. *Computer Speech & Language*, 24(3):433–444, 2010.

[ZC98] George Zavaliagkos and Thomas Colthurst. Utilizing Untranscribed Training Data to Improve Performance. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 301–305, 1998.

[Zho05] Shi Zhong. Semi-supervised Sequence Classification with HMMs. *International Journal of Pattern Recognition & Artificial Intelligence*, pages 165–182, 2005.

List of Notations

$[a, b]$	Closed interval $\{x \in \mathbb{R} : a \leq x \leq b\}$
$[n]$	The set $\{1, \ldots, n\}$ for any $n \geq 1$
2^A	The power set of A for any set A
$\boldsymbol{A}_P^{\varepsilon,m}$	Payoff matrix modelling PAC-learning under the fixed distribution P
$\boldsymbol{b}$	Vector of binary labels; $\boldsymbol{b} = c^*(\boldsymbol{x}) \in \{0, 1\}^m$
$\mathcal{C}$	Concept class
$\mathcal{C}^+$	All unions of concepts from $\mathcal{C}$
$\mathcal{C}^-$	All intersections of concepts from $\mathcal{C}$
c	Concept from $\mathcal{C}$
c^*	Target concept from $\mathcal{C}$
$\mathrm{co}(\mathcal{C})$	Complement of class $\mathcal{C}$; $\mathrm{co}(\mathcal{C}) = \{X \setminus c : c \in \mathcal{C}\}$
δ	Confidence parameter in PAC-learning
Δ	The d-dimensional probability simplex
$DIS(V)$	Disagreement region of the concepts in V
ε	Accuracy parameter in PAC-learning
$\mathbb{E}[A]$	Expected value of random variable A
$\mathcal{H}$	Hypothesis class
h	Hypothesis from $\mathcal{H}$
$\mathbb{I}(S)$	Indicator function of condition S; $\mathbb{I}(S) = 1$ if S holds, $\mathbb{I}(S) = 0$ otherwise
L	A learning algorithm
$\mathcal{L}$	A proper learning function; $\mathcal{L} : X^m \times \{0, 1\}^m \to \mathcal{C}$
$\ln(x)$	The natural logarithm of x
$\log_b(x)$	The logarithm of x to basis b
m	Usually the sample size

M	Number of learning functions
$m_{\mathcal{C}}^{L}$	Sample complexity of learner L and concept class $\mathcal{C}$
$m_{\mathcal{C},\mathcal{P}}^{L}$	Sample complexity of learner L and concept class $\mathcal{C}$ under the fixed distribution P
$m_{\mathcal{C}}^{*}$	Sample complexity of the best learner for concept class $\mathcal{C}$
$m_{\mathcal{C},\mathcal{P}}^{*}$	Sample complexity of the best learner L for concept class $\mathcal{C}$ under the fixed distribution P
N	Number of concepts
$N_{\mathcal{C},P}$	Covering number of class $\mathcal{C}$ under distribution P
$s(\mathcal{C})$	Singleton-size of $\mathcal{C}$; $s(\mathcal{C}) = 4 \cdot \max\{s^{+}(\mathcal{C}), s^{-}(\mathcal{C})\}$
$s^{+}(\mathcal{C})$	Size of the largest singleton subclass in class $\mathcal{C}$
$s^{-}(\mathcal{C})$	Size of the largest co-singleton subclass in class $\mathcal{C}$ (the “unique negative dimension” [GG89])
SF_n	Sunflower class; $\mathrm{SF}_n = \{\{0\}, \{0,1\}, \{0,2\}, \ldots, \{0,n\}\}$
θ	Disagreement coefficient
$\boldsymbol{p}$	Probability vector over all proper learning functions; corresponds to a mixed strategy for the learner
P	Domain distribution
$\mathcal{P}$	Family of domain distributions
$P(A)$	Probability of event A under domain distribution P
$\Pr(A)$	Probability of event A
$\boldsymbol{q}$	Probability vector over all concepts; corresponds to a mixed strategy for the opponent of the learner
$\boldsymbol{R}$	Payoff matrix; models PAC-learning in a worst-case analysis
$\bar{\boldsymbol{R}}$	Payoff matrix; models PAC-learning in an average-case analysis
$\mathrm{VCdim}(\mathcal{C})$	VC-dimension of class $\mathcal{C}$
$V_{\mathcal{H}}$	Version space; the set of all hypotheses in $\mathcal{H}$ consistent with the sample
$\boldsymbol{x}$	Vector of random sample points; $\boldsymbol{x} \sim P^{m}$
$(\boldsymbol{x}, \boldsymbol{b})$	Labeled sample
X	A domain

Note that this list contains the most frequently used notations only.

List of Figures

List of Tables

List of Algorithms