A Discriminative Model for Semi-Supervised Learning *

Maria-Florina Balcan and Avrim Blum
Computer Science Department
Carnegie Mellon University.
{ninamf,avrim}@cs.cmu.edu

Supervised learning — that is, learning from labeled examples — is an area of Machine Learning that has reached substantial maturity. It has generated general-purpose and practically-successful algorithms and the foundations are quite well understood and captured by theoretical frameworks such as the PAC-learning model and the Statistical Learning theory framework. However, for many contemporary practical problems such as classifying web pages or detecting spam, there is often additional information available in the form of *unlabeled* data, which is often much cheaper and more plentiful than labeled data. As a consequence, there has recently been substantial interest in *semi-supervised* learning — using unlabeled data together with labeled data — since any useful information that reduces the amount of labeled data needed can be a significant benefit. Several techniques have been developed for doing this, along with experimental results on a variety of different learning problems. Unfortunately, the standard learning frameworks for reasoning about supervised learning do not capture the key aspects and the assumptions underlying these *semi*-supervised learning methods.

In this paper we describe an augmented version of the PAC model designed for semi-supervised learning, that can be used to reason about many of the different approaches taken over the past decade in the Machine Learning community. This model provides a unified framework for analyzing when and why unlabeled data can help, in which one can analyze both sample-complexity and algorithmic issues. The model can be viewed as an extension of the standard PAC model where, in addition to a concept class $\mathcal{C}$, one also proposes a compatibility notion: a type of compatibility that one believes the target concept should have with the underlying distribution of data. Unlabeled data is then potentially helpful in this setting because it allows one to estimate compatibility over the space of hypotheses, and to reduce the size of the search space from the whole set of hypotheses $\mathcal{C}$ down to those that, according to one's assumptions, are a-priori reasonable with respect to the distribution. As we show, many of the assumptions underlying existing semi-supervised learning algorithms can be formulated in this framework.

After proposing the model, we then analyze sample-complexity issues in this setting: that is, how much of each type of data one should expect to need in order to learn well, and what the key quantities are that these numbers depend on. We also consider the algorithmic question of how to efficiently optimize for natural classes and compatibility notions, and provide several algorithmic results including an improved bound for Co-Training with linear separators when the distribution satisfies independence given the label.

Categories and Subject Descriptors: I.5.1 [**Pattern Recognition**]: Models; F.2.0 [**Analysis**

* A preliminary version of this paper appears as "A PAC-style Model for Learning from Labeled and Unlabeled Data", Proceedings of the 18th Annual Conference on Learning Theory (COLT), pp. 111 – 1264, 2005, and portions appear as Chapter 22 of the book *Semi-Supervised Learning* [Chapelle et al. 2006].

of Algorithms and Problem Complexity]: General; G.3 [**Probability and Statistics**]: Correlation and regression analysis

General Terms: Algorithms, Theory.

Additional Key Words and Phrases: Machine Learning, Semi-Supervised Learning, Value of Unlabeled Data, Sample Complexity, Cover Bounds, Uniform Convergence Bounds, Structural Risk Minimization, Data Dependent SRM, Efficient Learning Algorithms, Multi-view Classification.

1. INTRODUCTION

In recent years there has been substantial and growing interest in using unlabeled data together with labeled data in machine learning. The motivation is clear: in many applications, unlabeled data can be much cheaper and much more plentiful than labeled data. If useful information can be extracted from unlabeled examples that allows for learning from fewer labeled examples, this can be a substantial benefit. A number of Semi-Supervised learning techniques have been developed for doing this, along with experimental results on a variety of different learning problems. These include label propagation for word-sense disambiguation [Yarowsky 1995], co-training for classifying web pages [Blum and Mitchell 1998] and improving visual detectors [Levin et al. 2003], transductive SVM [Joachims 1999] and EM [Nigam et al. 2000] for text classification, graph-based methods [Zhu et al. 2003c], and others. The problem of learning from labeled and unlabeled data has been the topic of several ICML workshops [Ghani et al. 2003; Amini et al. 2005] as well as a recent book [Chapelle et al. 2006] and survey article [Zhu 2006].

What makes unlabeled data so useful and what many of these methods exploit, is that for a wide variety of learning problems, the natural regularities of the problem involve not only the *form* of the function being learned by also how this function *relates* to the distribution of data. For example, in many problems one might expect the target function should cut through low density regions of the space, a property used by the transductive SVM algorithm [Joachims 1999]. In other problems one might expect the target to be self-consistent in some way, a property used by Co-training [Blum and Mitchell 1998]. Unlabeled data is potentially useful in these settings because it then allows one to reduce the search space to a set which is a-priori reasonable with respect to the underlying distribution.

Unfortunately, however, the underlying assumptions of these semi-supervised learning methods are not captured well by standard theoretical models. The main goal of this work is to propose a *unified theoretical framework* for semi-supervised learning, in which one can analyze when and why unlabeled data can help, and in which one can discuss both sample-complexity and algorithmic issues in a discriminative (PAC-model style) framework.

One difficulty from a theoretical point of view is that standard discriminative learning models do not allow one to specify relations that one believes the target should have with the underlying distribution. In particular, both in the PAC model [Valiant 1984; Blumer et al. 1989; Kearns and Vazirani 1994] and the Statistical Learning Theory framework [Vapnik 1998] there is purposefully a complete disconnect between the data distribution D and the target function f being learned.

The only prior belief is that f belongs to some class $\mathcal{C}$: even if the data distribution D is known fully, any function $f \in \mathcal{C}$ is still possible. For instance, in the PAC model, it is perfectly natural (and common) to talk about the problem of learning a concept class such as DNF formulas [Linial et al. 1989; Verbeurgt 1990] or an intersection of halfspaces [Baum 1990; Blum and Kannan 1997; Vempala 1997; Klivans et al. 2002] over the uniform distribution; but clearly in this case unlabeled data is useless — you can just generate it yourself. For learning over an unknown distribution, unlabeled data can help somewhat in the standard models (e.g., by allowing one to use distribution-specific algorithms and sample-complexity bounds [Benedek and Itai 1991; Kaariainen 2005]), but this does not seem to capture the power of unlabeled data in practical semi-supervised learning methods.

In *generative* models, one *can* easily talk theoretically about the use of unlabeled data, e.g., [Castelli and Cover 1995; 1996]. However, these results typically make strong assumptions that essentially imply that there is only one natural distinction to be made for a given (unlabeled) data distribution. For instance, a typical generative model would be that we assume positive examples are generated by one Gaussian, and negative examples are generated by another Gaussian. In this case, given enough unlabeled data, we could in principle recover the Gaussians and would need labeled data only to tell us which Gaussian is the positive one and which is the negative one.[1] However, this is too strong an assumption for most real-world settings. Instead, we would like our model to allow for a distribution over data (e.g., documents we want to classify) where there are a number of plausible distinctions we might want to make. In addition, we would like a general framework that can be used to model many different uses of unlabeled data.

1.1 Our Contribution

In this paper, we present a PAC-style framework that bridges between these positions and can be used to help think about and analyze many of the ways unlabeled data is typically used. This framework extends the PAC learning model in a way that allows one to express not only the form of target function one is considering, but also relationships that one hopes the target function and underlying distribution will possess. We then analyze both sample-complexity issues—that is, how much of each type of data one should expect to need in order to learn well—as well as algorithmic results in this model. We derive bounds for both the realizable (PAC) and agnostic (statistical learning framework) settings.

Specifically, the idea of the proposed model is to augment the PAC notion of a *concept class*, which is a set of functions (such as linear separators or decision trees), with a notion of *compatibility* between a function and the data distribution that we hope the target function will satisfy. Rather than talking of "learning a concept class $\mathcal{C}$," we will talk of "learning a concept class $\mathcal{C}$ under compatibility notion χ." For example, suppose we believe there should exist a low-error linear separator, and that furthermore, if the data happens to cluster, then this separator does not slice through the middle of any such clusters. Then we would want a compatibility notion that penalizes functions that do, in fact, slice through clusters.

[1] [Castelli and Cover 1995; 1996] do not assume Gaussians in particular, but they do assume the distributions are distinguishable, which from this perspective has the same issue.

In this framework, the ability of unlabeled data to help depends on two quantities: first, the extent to which the target function indeed satisfies the given assumptions, and second, the extent to which the distribution allows this assumption to rule out alternative hypotheses. For instance, if the data does not cluster at all (say the underlying distribution is uniform in a ball), then all functions would equally satisfy this compatibility notion and the assumption is not useful. From a Bayesian perspective, one can think of this as a PAC model for a setting in which one's prior is not just over functions, but also over how the function and underlying distribution relate to each other.

To make our model formal, we will need to ensure that the degree of compatibility be something that can be *estimated from a finite sample*. To do this, we will require that the compatibility notion χ in fact be a function from $\mathcal{C} \times X$ to $[0, 1]$, where the compatibility of a hypothesis h with the data distribution D is then $\mathbf{E}_{x \sim D}[\chi(h, x)]$. That is, we require that the degree of *in*compatibility be a kind of unlabeled loss function, and the incompatibility of a hypothesis h with a data distribution D is a quantity we can think of as an "unlabeled error rate" that measures how a-priori unreasonable we believe some proposed hypothesis to be. For instance, in the example above of a "margin-style" compatibility, we could define $\chi(f, x)$ to be an increasing function of the distance of x to the separator f. In this case, the unlabeled error rate, $1 - \chi(f, D)$, is a measure of the probability mass close to the proposed separator. In co-training, where each example x has two "views" ($x = \langle x_1, x_2 \rangle$), the underlying belief is that the true target c^* can be decomposed into functions $\langle c_1^*, c_2^* \rangle$ over each view such that for most examples, $c_1^*(x_1) = c_2^*(x_2)$. In this case, we can define $\chi(\langle f_1, f_2 \rangle, \langle x_1, x_2 \rangle) = 1$ if $f_1(x_1) = f_2(x_2)$, and 0 if $f_1(x_1) \neq f_2(x_2)$. Then the compatibility of a hypothesis $\langle f_1, f_2 \rangle$ with an underlying distribution D is $\mathbf{Pr}_{\langle x_1, x_2 \rangle \sim D}[f_1(x_1) = f_2(x_2)]$.

This framework allows us to analyze the ability of a finite unlabeled sample to reduce our dependence on labeled examples, as a function of (1) the compatibility of the target function (i.e., how correct we were in our assumption) and (2) various measures of the "helpfulness" of the distribution. In particular, in our model, we find that unlabeled data can help in several distinct ways.

- If the target function is highly compatible with D and belongs to $\mathcal{C}$, then if we have enough unlabeled data to estimate compatibility over all $f \in \mathcal{C}$, we can in principle reduce the size of the search space from $\mathcal{C}$ down to just those $f \in \mathcal{C}$ whose estimated compatibility is high. For instance, if D is "helpful", then the set of such functions will be much smaller than the entire set $\mathcal{C}$. In the agnostic case we can do (unlabeled)-data-dependent structural risk minimization to trade off labeled error and incompatibility.
- By providing an estimate of D, unlabeled data can allow us to use a more refined distribution-specific notion of "hypothesis space size" such as Annealed VC-entropy [Devroye et al. 1996], Rademacher complexities [Koltchinskii 2001; Bartlett and Mendelson 2002; Boucheron et al. 2005] or the size of the smallest ϵ-cover [Benedek and Itai 1991], rather than VC-dimension [Blumer et al. 1989; Kearns and Vazirani 1994]. In fact, for many natural notions of compatibility we find that the sense in which unlabeled data reduces the "size" of the search space is best described in these distribution-specific measures.

- Finally, if the distribution is especially helpful, we may find that not only does the set of compatible $f \in \mathcal{C}$ have a small ϵ-cover, but also the elements of the cover are far apart. In that case, if we assume the target function is fully compatible, we may be able to learn from even fewer labeled examples than the $\Omega(1/\epsilon)$ needed just to *verify* a good hypothesis. For instance, as one application of this, we show that under the assumption of independence given the label, one can efficiently perform Co-Training of linear separators from a single labeled example!

Our framework also allows us to address the issue of how much *unlabeled* data we should expect to need. Roughly, the "VCdim/ϵ^2" form of standard sample complexity bounds now becomes a bound on the number of *unlabeled* examples we need to uniformly estimate compatibilities. However, technically, the set whose VC-dimension we now care about is not $\mathcal{C}$ but rather a set defined by both $\mathcal{C}$ and χ: that is, the overall complexity depends both on the complexity of $\mathcal{C}$ and the complexity of the notion of compatibility (see Section 3.1.2). One consequence of our model is that if the target function and data distribution are both well behaved with respect to the compatibility notion, then the sample-size bounds we get for labeled data can substantially beat what one could hope to achieve through pure labeled-data bounds, and we illustrate this with a number of examples through the paper.

1.2 Summary of Main Results

The primary contributions of this paper are three-fold. First, as described above, we develop a new discriminative (PAC-style) model for semi-supervised learning, that can be used to analyze when unlabeled data can help and how *much* unlabeled data is needed in order to gain its benefits, as well as the algorithmic problems involved. Second, we present a number of sample-complexity bounds in this framework, both in terms of uniform-convergence results—which apply to any algorithm that is able to find rules of low error and high compatibility—as well as ϵ-cover-based bounds that apply to a more restricted class of algorithms but can be substantially tighter. For instance, we describe several natural cases in which ϵ-cover-based bounds can apply even though with high probability there still exist bad hypotheses in the class consistent with the labeled and unlabeled examples. Finally, we present several PAC-style algorithmic results in this model. Our main algorithmic result is a new algorithm for Co-Training with linear separators that, if the distribution satisfies independence given the label, requires only a single labeled example to learn to any desired error rate ϵ *and* is computationally efficient (i.e., achieves PAC guarantees). This substantially improves on the results of [Blum and Mitchell 1998] which required enough labeled examples to produce an initial weak hypothesis, and in the process we get a simplification to the noisy halfspace learning algorithm of [Blum et al. 1998].

Our framework has helped analyze many of the existing semi-supervised learning methods used in practice and has guided the development of new semi-supervised learning algorithms and analyses. We discuss this further in Section 6.1.

1.3 Structure of this Paper

We begin by describing the general setting in which our results apply as well as several examples to illustrate our framework in Section 2. We then give results both for *sample complexity* (in principle, how much data is needed to learn) and *efficient algorithms*. In terms of sample-complexity, we start by discussing uniform convergence results in Section 3.1. For clarity we begin with the case of finite hypothesis spaces in Section 3.1.1, and then discuss infinite hypothesis spaces in Section 3.1.2. These results give bounds on the number of examples needed for any learning algorithm that produces a compatible hypothesis of low empirical error. We also show how in the agnostic case we can do (unlabeled)-data-dependent structural risk minimization to trade off labeled error and incompatibility in Section 3.1.3. To achieve tighter bounds, in Section 3.2 we give results based on the notion of ϵ-cover size. These bounds hold only for algorithms of a specific type (that first use the unlabeled data to choose a small set of "representative" hypotheses and then choose among the representatives based on the labeled data), but can yield bounds substantially better than with uniform convergence (e.g., we can learn even though there exist bad $h \in \mathcal{C}$ consistent with the labeled and unlabeled examples).

In Section 4, we give our algorithmic results. We begin with a particularly simple class $\mathcal{C}$ and compatibility notion χ for illustration, and then give our main algorithmic result for Co-Training with linear separators. In Section 5 we discuss a transductive analog of our model, connections with generative models and other ways of using unlabeled data in machine learning, as well as the relationship between our model and the Luckiness Framework [Shawe-Taylor et al. 1998] developed in the context of supervised learning. Finally, in Section 6 we discuss some implications of our model and present our conclusions, as well a number of open problems.

2. A FORMAL FRAMEWORK

In this section we introduce general notation and terminology we use throughout the paper, and describe our model for semi-supervised learning. In particular, we formally define what we mean by a *notion of compatibility* and we illustrate it through a number of examples including margins and co-training.

We will focus on binary classification problems. We assume that our data comes according to a fixed unknown distribution D over an instance space $\mathcal{X}$, and is labeled by some unknown target function $c^* : \mathcal{X} \to \{0, 1\}$. A learning algorithm is given a set S_L of labeled examples drawn i.i.d. from D and labeled by c^* as well as a (usually larger) set S_U of unlabeled examples from D. The goal is to perform some optimization over the samples S_L and S_U and to output a hypothesis that agrees with the target over most of the distribution. In particular, the error rate (also called "0-1 loss") of a given hypothesis f is defined as $err(f) = err_D(f) = \mathbf{Pr}_{x\sim D}[f(x) \neq c^*(x)]$. For any two hypotheses f_1, f_2, the distance with respect to D between f_1 and f_2 is defined as $d(f_1, f_2) = d_D(f_1, f_2) = \mathbf{Pr}_{x\sim D}[f_1(x) \neq f_2(x)]$. We will use $\widehat{err}(f)$ to denote the empirical error rate of f on a given labeled sample (i.e., the fraction of mistakes on the sample) and $\hat{d}(f_1, f_2)$ to denote the empirical distance between f_1 and f_2 on a given unlabeled sample (the fraction of the sample on which they disagree). As in the standard PAC model, a *concept class* or *hypothesis space* is a set of functions over the instance space $\mathcal{X}$. In the "realizable case", we make

the assumption that the target is in a given class $\mathcal{C}$, whereas in the "agnostic case" we do not make this assumption and instead aim to compete with the best function in the given class $\mathcal{C}$.

We now formally describe what we mean by a notion of compatibility. A *notion of compatibility* is a mapping from a hypothesis f and a distribution D to $[0,1]$ indicating how "compatible" f is with D. In order for this to be estimable from a finite sample, we require that compatibility be an expectation over individual examples.[2] Specifically, we define:

DEFINITION 1. *A* legal notion of compatibility *is a function $\chi : \mathcal{C} \times \mathcal{X} \to [0,1]$ where we (overloading notation) define $\chi(f,D) = \mathbf{E}_{x\sim D}[\chi(f,x)]$. Given a sample S, we define $\chi(f,S)$ to be the empirical average of χ over the sample.*

NOTE 1. *One could also allow compatibility functions over k-tuples of examples, in which case our (unlabeled) sample-complexity bounds would simply increase by a factor of k. For settings in which D is actually known in advance (e.g., transductive learning, see Section 5.1) we can drop this requirement entirely and allow any notion of compatibility $\chi(f,D)$ to be legal.*

DEFINITION 2. *Given compatibility notion χ, the* incompatibility *of f with D is $1-\chi(f,D)$. We will also call this its* unlabeled error rate*, $err_{unl}(f)$, when χ and D are clear from context. For a given sample S, we use $\widehat{err}_{unl}(f) = 1-\chi(f,S)$ to denote the empirical average over S.*

Finally, we need a notation for the set of functions whose incompatibility is at most some given value τ.

DEFINITION 3. *Given value τ, we define $\mathcal{C}_{D,\chi}(\tau) = \{f \in \mathcal{C} : err_{unl}(f) \leq \tau\}$. So, e.g., $\mathcal{C}_{D,\chi}(1) = \mathcal{C}$. Similarly, for a sample S, we define $\mathcal{C}_{S,\chi}(\tau) = \{f \in \mathcal{C} : \widehat{err}_{unl}(f) \leq \tau\}$*

We now give several examples to illustrate this framework:

Example 1. Suppose examples are points in R^d and $\mathcal{C}$ is the class of linear separators. A natural belief in this setting is that data should be "well-separated": not only should the target function separate the positive and negative examples, but it should do so by some reasonable *margin* γ. This is the assumption used by Transductive SVM, also called Semi-Supervised SVM (S^3VM) [Joachims 1999; Bie and Cristianini 2003; Chapelle and Zien 2005]. In this case, if we are given γ up front, we could define $\chi(f,x) = 1$ if x is farther than distance γ from the hyperplane defined by f, and $\chi(f,x) = 0$ otherwise. So, the incompatibility of f with D is the *probability mass within distance* γ of the hyperplane $f \cdot x = 0$. Alternatively, if we do not want to commit to a specific γ in advance, we could define $\chi(f,x)$ to be a smooth function of the distance of x to the separator, as

[2]One could imagine more general notions of compatibility with the property that they can be estimated from a finite sample and all our results would go through in that case as well. We consider the special case where the compatibility is an expectation over individual examples for simplicity of notation, and because most existing semi-supervised learning algorithms used in practice do satisfy it.

done in [Chapelle and Zien 2005]. Note that in contrast, defining compatibility of a hypothesis based on the largest γ such that D has probability mass *exactly zero* within distance γ of the separator would *not* fit our model: it cannot be written as an expectation over individual examples and indeed would not be a good definition since one cannot distinguish "zero" from "exponentially close to zero" from a small sample of unlabeled data.

Example 2. In co-training [Blum and Mitchell 1998], we assume examples x each contain two "views": $x = \langle x_1, x_2 \rangle$, and our goal is to learn a pair of functions $\langle f_1, f_2 \rangle$, one on each view. For instance, if our goal is to classify web pages, we might use x_1 to represent the words on the page itself and x_2 to represent the words attached to links pointing *to* this page from other pages. The hope underlying co-training is that the two parts of the example are generally consistent, which then allows the algorithm to bootstrap from unlabeled data. For example, *iterative co-training* uses a small amount of labeled data to learn some initial information (e.g., if a link with the words "my advisor" points to a page then that page is probably a faculty member's home page). Then, when it finds an unlabeled example where one side is confident (e.g., the link says "my advisor"), it uses that to label the example for training over the other view. In *regularized co-training*, one attempts to directly optimize a weighted combination of accuracy on labeled data and agreement over unlabeled data. These approaches have been used for a variety of learning problems, including named entity classification [Collins and Singer 1999], text classification [Nigam and Ghani 2000; Ghani 2001], natural language processing [Pierce and Cardie 2001], large scale document classification [Park and Zhang 2003], and visual detectors [Levin et al. 2003]. As mentioned in Section 1, the assumptions underlying this method fit naturally into our framework. In particular, we can define the incompatibility of some hypothesis $\langle f_1, f_2 \rangle$ with distribution D as $\mathbf{Pr}_{\langle x_1, x_2 \rangle \sim D}[f_1(x_1) \neq f_2(x_2)]$. Similar notions are given in subsequent work of [Rosenberg and Bartlett 2007; Sridharan and Kakade 2008] for other types of learning problems (e.g. regression) and for other loss functions.

Example 3. In transductive graph-based methods, we are given a set of unlabeled examples connected in a graph G, where the interpretation of an edge is that we believe the two endpoints of the edge should have the *same* label. Given a few labeled vertices, various graph-based methods then attempt to use them to infer labels for the remaining points. If we are willing to view D as a distribution over *edges* (a uniform distribution if G is unweighted), then as in co-training we can define the incompatibility of some hypothesis f as the probability mass of edges that are cut by f, which then motivates various cut-based algorithms. For instance, if we require f to be boolean, then the mincut method of [Blum and Chawla 2001] finds the most-compatible hypothesis consistent with the labeled data; if we allow f to be fractional and define $1 - \chi(f, \langle x_1, x_2 \rangle) = (f(x_1) - f(x_2))^2$, then the algorithm of [Zhu et al. 2003c] finds the most-compatible consistent hypothesis. If we do not wish to view D as a distribution over edges, we could have D be a distribution over *vertices* and broaden Definition 1 to allow for χ to be a function over *pairs* of examples. In fact, as mentioned in Note 1, since we have perfect knowledge of D in this setting we can allow any compatibility function $\chi(f, D)$ to be legal. We discuss more connections with graph-based methods in Section 5.1.

Example 4. As a special case of co-training, suppose examples are pairs of points in R^d, $\mathcal{C}$ is the class of linear separators, and we believe the two points in each pair should both be on the *same* side of the target function. (So, this is a version of co-training where we require $f_1 = f_2$.) The motivation is that we want to use pairwise information as in Example 3, but we also want to use the features of each data point. For instance, in the word-sense disambiguation problem studied by [Yarowsky 1995], the goal is to determine which of several dictionary definitions is intended for some target word in a piece of text (e.g., is "plant" being used to indicate a tree or a factory?). The local context around each word can be viewed as placing it into R^d, but the edges correspond to a completely different type of information: the belief that if a word appears twice in the same document, it is probably being used in the *same* sense both times. In this setting, we could use the same compatibility function as in Example 3, but rather than having the concept class $\mathcal{C}$ be all possible functions, we restrict $\mathcal{C}$ to just linear separators.

Example 5**.** In a related setting to co-training considered by [Leskes 2005], examples are single points in $\mathcal{X}$ but we have a pair of hypothesis spaces $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$ (or more generally a k-tuple $\langle \mathcal{C}_1, \ldots, \mathcal{C}_k \rangle$), and the goal is to find a pair of hypotheses $\langle f_1, f_2 \rangle \in \mathcal{C}_1 \times \mathcal{C}_2$ with low error over labeled data and that agree over the distribution. For instance, if data is sufficiently "well-separated", one might expect there to exist both a good linear separator and a good decision tree, and one would like to use this assumption to reduce the need for labeled data. In this case one could define compatibility of $\langle f_1, f_2 \rangle$ with D as $\mathbf{Pr}_{x \sim D}[f_1(x) = f_2(x)]$, or the similar notions given in [Leskes 2005; Shawe-Taylor 2006].

3. SAMPLE COMPLEXITY RESULTS

We now present several sample-complexity bounds that can be derived in this framework, showing how unlabeled data, together with a suitable compatibility notion, can reduce the need for labeled examples. We do not focus on giving the tightest possible bounds, but instead on the types of bounds and the quantities on which they depend, in order to better understand what it is about the learning *problem* one can hope to leverage from with unlabeled data.

The high-level structure of all of these results is as follows. First, given enough unlabeled data (where "enough" will be a function of some measure of the complexity of $\mathcal{C}$ and possibly of χ as well), we can uniformly estimate the true compatibilities of all functions in $\mathcal{C}$ using their empirical compatibilities over the sample. Then, by using this quantity to give a preference ordering over the functions in $\mathcal{C}$, in the realizable case we can reduce "$\mathcal{C}$" down to "the set of functions in $\mathcal{C}$ whose compatibility is not much larger than the true target function" in bounds for the number of *labeled* examples needed for learning. In the agnostic case we can do (unlabeled)-data-dependent structural risk minimization to trade off labeled error and incompatibility. The specific bounds differ in terms of the exact complexity measures used (and a few other issues) and we provide examples illustrating when and how certain complexity measures can be significantly more powerful than others. Moreover, one can prove fallback properties of these procedures — the number of labeled examples required is never much worse than the number of labeled examples required by a standard supervised learning algorithm. However, if the as-

sumptions happen to be right, one can significantly benefit by using the unlabeled data.

3.1 Uniform Convergence Bounds

We begin with uniform convergence bounds (later in Section 3.2 we give tighter ϵ-cover bounds that apply to algorithms of a particular form). For clarity, we begin with the case of finite hypothesis spaces where we measure the "size" of a set of functions by just the number of functions in the set. We then discuss several issues that arise when considering infinite hypothesis spaces, such as what is an appropriate measure for the "size" of the set of compatible functions, and the need to account for the complexity of the compatibility notion itself. Note that in the standard PAC model, one typically talks of either the realizable case, where we assume that the target function c^* belongs to $\mathcal{C}$, or the agnostic case where we allow any target function c^* [Kearns and Vazirani 1994]. In our setting, we have the additional issue of *unlabeled* error rate, and can either make an a-priori assumption that the target function's unlabeled error is low, or else provide a bound in which our sample size (or error rate) depends on whatever its unlabeled error happens to be. We begin in Sections 3.1.1 and 3.1.2 with bounds for the the setting in which we assume $c^* \in \mathcal{C}$, and then in Section 3.1.3 we consider the agnostic case where we remove this assumption.

3.1.1 *Finite hypothesis spaces.* We first give a bound for the "doubly realizable" case where we assume $c^* \in \mathcal{C}$ and $err_{unl}(c^*) = 0$.

THEOREM 4. *If $c^* \in \mathcal{C}$ and $err_{unl}(c^*) = 0$, then m_u unlabeled examples and m_l labeled examples are sufficient to learn to error ϵ with probability $1 - \delta$, where*

$$m_u = \frac{1}{\epsilon}\left[\ln|\mathcal{C}| + \ln\frac{2}{\delta}\right] \quad \textit{and} \quad m_l = \frac{1}{\epsilon}\left[\ln|\mathcal{C}_{D,\chi}(\epsilon)| + \ln\frac{2}{\delta}\right].$$

In particular, with probability at least $1 - \delta$, all $f \in \mathcal{C}$ with $\widehat{err}(f) = 0$ and $\widehat{err}_{unl}(f) = 0$ have $err(f) \leq \epsilon$.

PROOF. The probability that a given hypothesis f with $err_{unl}(f) > \epsilon$ has $\widehat{err}_{unl}(f) = 0$ is at most $(1-\epsilon)^{m_u} < \frac{\delta}{2|\mathcal{C}|}$ for the given value of m_u. Therefore, by the union bound, the number of unlabeled examples is sufficient to ensure that with probability $1 - \frac{\delta}{2}$, only hypotheses in $\mathcal{C}_{D,\chi}(\epsilon)$ have $\widehat{err}_{unl}(f) = 0$. The number of labeled examples then similarly ensures that with probability $1 - \frac{\delta}{2}$, none of those whose true error is at least ϵ have an empirical error of 0, yielding the theorem. □

Interpretation: If the target function indeed is perfectly correct and compatible, then Theorem 4 gives sufficient conditions on the number of examples needed to ensure that an algorithm that optimizes both quantities over the observed data will, in fact, achieve a PAC guarantee. To emphasize this, we will say that an algorithm efficiently PAC_{SSL}-learns the pair $(\mathcal{C}, \chi)$ if it is able to achieve a PAC guarantee using time and sample sizes polynomial in the bounds of Theorem 4. For a formal definition see Definition 7 at the end of this section.

We can think of Theorem 4 as bounding the number of labeled examples we need as a function of the "helpfulness" of the distribution D with respect to our

notion of compatibility. That is, in our context, a helpful distribution is one in which $\mathcal{C}_{D,\chi}(\epsilon)$ is small, and so we do not need much labeled data to identify a good function among them. We can get a similar bound in the situation when the target function is not fully compatible:

THEOREM 5. *If $c^* \in \mathcal{C}$ and $err_{unl}(c^*) = t$, then m_u unlabeled examples and m_l labeled examples are sufficient to learn to error ϵ with probability $1-\delta$, for*

$$m_u = \frac{2}{\epsilon^2}\left[\ln|\mathcal{C}| + \ln\frac{4}{\delta}\right] \quad \textit{and} \quad m_l = \frac{1}{\epsilon}\left[\ln|\mathcal{C}_{D,\chi}(t+2\epsilon)| + \ln\frac{2}{\delta}\right].$$

In particular, with probability at least $1-\delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$.

Alternatively, given the above number of unlabeled examples m_u, for any *number of labeled examples m_l, with probability at least $1-\delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has*

$$err(f) \;\leq\; \frac{1}{m_l}\left[\ln|\mathcal{C}_{D,\chi}(err_{unl}(c^*)+2\epsilon)| + \ln\frac{2}{\delta}\right]. \tag{1}$$

PROOF. By Hoeffding bounds, m_u is sufficiently large so that with probability at least $1-\delta/2$, all $f \in \mathcal{C}$ have $|\widehat{err}_{unl}(f) - err_{unl}(f)| \leq \epsilon$. Thus, $\{f \in \mathcal{C} : \widehat{err}_{unl}(f) \leq t+\epsilon\} \subseteq \mathcal{C}_{D,\chi}(t+2\epsilon)$. For the first implication, the given bound on m_l is sufficient so that with probability at least $1-\delta$, all $f \in \mathcal{C}$ with $\widehat{err}(f) = 0$ and $\widehat{err}_{unl}(f) \leq t+\epsilon$ have $err(f) \leq \epsilon$; furthermore, $\widehat{err}_{unl}(c^*) \leq t+\epsilon$, so such a function f exists. Therefore, with probability at least $1-\delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$, as desired. For second implication, inequality (1) follows immediately by solving for the labeled estimation-error as a function of m_l. □

Interpretation: Theorem 5 has several implications. Specifically:

(1) If we can optimize the (empirical) unlabeled error rate subject to having zero empirical labeled error, then to achieve low true error it suffices to draw a number of labeled examples that depends logarithmically on the number of functions in $\mathcal{C}$ whose unlabeled error rate is at most 2ϵ greater than that of the target c^*.

(2) Alternatively, for *any* given number of labeled examples m_l, we can provide a bound (given in equation 1) on our error rate that again depends logarithmically on the number of such functions.

(3) If we have a desired maximum error rate ϵ and do not know the value of $err_{unl}(c^*)$ but have the ability to draw additional labeled examples as needed, then we can simply do a standard "doubling trick" on m_l. On each round, we check if the hypothesis f found indeed has sufficiently low empirical unlabeled error rate, and we spread the "δ" parameter across the different runs. See, e.g., Corollary 10 in Section 3.1.2.

Finally, before going to infinite hypothesis spaces, we give a simple Occam-style version of the above bounds for this setting. Given a sample S, let us define $\mathsf{desc}_S(f) = \ln|\mathcal{C}_{S,\chi}(\widehat{err}_{unl}(f))|$. That is, $\mathsf{desc}_S(f)$ is the description length of f

(in "nats") if we sort hypotheses by their empirical compatibility and output the index of f in this ordering. Similarly, define ϵ-$\mathsf{desc}_D(f) = \ln|\mathcal{C}_{D,\chi}(err_{unl}(f)+\epsilon)|$. This is an upper-bound on the description length of f if we sort hypotheses by an ϵ-approximation to the their true compatibility. Then we immediately get a bound as follows:

COROLLARY 6. *For any set S of unlabeled data, given m_l labeled examples, with probability at least $1-\delta$,* all $f \in \mathcal{C}$ *satisfying $\widehat{err}(f) = 0$ and $\mathsf{desc}_S(f) \leq \epsilon m_l - \ln(1/\delta)$ have $err(f) \leq \epsilon$. Furthermore, if $|S| \geq \frac{2}{\epsilon^2}[\ln|\mathcal{C}| + \ln\frac{2}{\delta}]$, then with probability at least $1-\delta$, all $f \in \mathcal{C}$ satisfy $\mathsf{desc}_S(f) \leq \epsilon$-$\mathsf{desc}_D(f)$.*

Interpretation: The point of this bound is that an algorithm can use observable quantities (the "empirical description length" of the hypothesis produced) to determine if it can be confident that its true error rate is low. Furthermore, if we have enough unlabeled data, the observable quantities will be no worse than if we were learning a slightly less compatible function using an infinite-size unlabeled sample.

Note that if we begin with a non-distribution-dependent ordering of hypotheses, inducing some description length $\mathsf{desc}(f)$, and our compatibility assumptions turn out to be wrong, then it could well be that $\mathsf{desc}_D(c^*) > \mathsf{desc}(c^*)$. In this case our use of unlabeled data would end up hurting rather than helping. However, notice that by merely interleaving the initial ordering and the ordering produced by S, we get a new description length $\mathsf{desc}_{new}(f)$ such that $\mathsf{desc}_{new}(f) \leq 1 + \min(\mathsf{desc}(f), \mathsf{desc}_S(f))$. Thus, up to an additive constant, we can get the best of both orderings.

Also, if we have the ability to purchase additional labeled examples until the function produced is sufficiently "short" compared to the amount of data, then we can perform the usual stratification and be confident whenever we find a consistent function f such that $\mathsf{desc}_S(f) \leq \epsilon m_l - \ln(\frac{m_l(m_l+1)}{\delta})$, where m_l is the number of labeled examples seen so far.

Efficient algorithms in our model Finally, we end this section with a definition describing our goals for efficient learning algorithms, based on the above sample bounds.

DEFINITION 7. *Given a class $\mathcal{C}$ and compatibility notion χ, we say that an algorithm efficiently* $\mathrm{PAC}_{\mathrm{SSL}}$*-learns the pair $(\mathcal{C},\chi)$ if, for any distribution D, for any target function $c^* \in \mathcal{C}$ with $err_{unl}(c^*) = 0$, for any given $\epsilon,\delta > 0$, with probability at least $1-\delta$ it achieves error at most ϵ using poly$(\log|C|, 1/\epsilon, 1/\delta)$ unlabeled examples and poly$(\log|C_{D,\chi}(\epsilon)|, 1/\epsilon, 1/\delta)$ labeled examples, and with time poly$(\log|C|, 1/\epsilon, 1/\delta)$.*

We say that an algorithm semi-agnostically $\mathrm{PAC}_{\mathrm{SSL}}$*-learns $(\mathcal{C},\chi)$ if it is able to achieve this guarantee for any $c^* \in \mathcal{C}$ even if $err_{unl}(c^*) \neq 0$, using labeled examples poly$(\log|C_{D,\chi}(err_{unl}(c^*)+\epsilon)|, 1/\epsilon, 1/\delta)$.*

3.1.2 *Infinite hypothesis spaces.* To reduce notation, we will assume in the rest of this paper that $\chi(f,x) \in \{0,1\}$ so that $\chi(f,D) = \mathbf{Pr}_{x\sim D}[\chi(f,x) = 1]$. However, all our sample complexity results can be easily extended to the general case.

For infinite hypothesis spaces, the first issue that arises is that in order to achieve uniform convergence of *unlabeled* error rates, the set whose complexity we care about is not $\mathcal{C}$ but rather $\chi(\mathcal{C}) = \{\chi_f : f \in \mathcal{C}\}$ where we define $\chi_f(x) = \chi(f,x)$. For

instance, suppose examples are just points on the line, and $\mathcal{C} = \{f_a(x) : f_a(x) = 1 \text{ iff } x \leq a\}$. In this case, $\text{VCdim}(\mathcal{C}) = 1$. However, we could imagine a compatibility function such that $\chi(f_a, x)$ depends on some complicated relationship between the real numbers a and x. In this case, $\text{VCdim}(\chi(\mathcal{C}))$ is much larger, and indeed we would need many more unlabeled examples to estimate compatibility over all of $\mathcal{C}$.

A second issue is that we need an appropriate measure for the "size" of the set of surviving functions. VC-dimension tends not to be a good choice: for instance, if we consider the case of Example 1 (margins), then even if data is concentrated in two well-separated "blobs", the set of compatible separators still has as large a VC-dimension as the entire class even though they are all very similar with respect to D (see, e.g., Figure 1 after Theorem 9 below). Instead, it is better to consider distribution dependent complexity measures such as annealed VC-entropy [Devroye et al. 1996] or Rademacher averages [Koltchinskii 2001; Bartlett and Mendelson 2002; Boucheron et al. 2005]. For this we introduce some notation. Specifically, for any $\mathcal{C}$, we denote by $\mathcal{C}[m, D]$ the expected number of splits of m points (drawn i.i.d.) from D using concepts in $\mathcal{C}$. Also, for a given (fixed) $S \subseteq \mathcal{X}$, we will denote by $\overline{S}$ the uniform distribution over S, and by $\mathcal{C}[m, \overline{S}]$ the expected number of splits of m points from $\overline{S}$ using concepts in $\mathcal{C}$. The following is the analog of Theorem 5 for the infinite case.

THEOREM 8. *If $c^* \in \mathcal{C}$ and $err_{unl}(c^*) = t$, then m_u unlabeled examples and m_l labeled examples are sufficient to learn to error ϵ with probability $1 - \delta$, for*

$$m_u = \mathcal{O}\left(\frac{VCdim\left(\chi(\mathcal{C})\right)}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{2}{\delta}\right)$$

and

$$m_l = \frac{2}{\epsilon}\left[\ln\left(2\mathcal{C}_{D,\chi}(t + 2\epsilon)[2m_l, D]\right) + \ln \frac{4}{\delta}\right],$$

where recall $\mathcal{C}_{D,\chi}(t+2\epsilon)[2m_l, D]$ is the expected number of splits of $2m_l$ points drawn from D using concepts in $\mathcal{C}$ of unlabeled error rate $\leq t + 2\epsilon$. In particular, with probability at least $1 - \delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$.

PROOF. Let S be the set of m_u unlabeled examples. By standard VC-dimension bounds (e.g., see Theorem 20 in Appendix A) the number of unlabeled examples given is sufficient to ensure that with probability at least $1 - \frac{\delta}{2}$ we have $|\mathbf{Pr}_{x\sim S}[\chi_f(x) = 1] - \mathbf{Pr}_{x\sim D}[\chi_f(x) = 1]| \leq \epsilon$ for all $\chi_f \in \chi(\mathcal{C})$. Since $\chi_f(x) = \chi(f, x)$, this implies that we have $|\widehat{err}_{unl}(f) - err_{unl}(f)| \leq \epsilon$ for all $f \in \mathcal{C}$. So, the set of hypotheses with $\widehat{err}_{unl}(f) \leq t + \epsilon$ is contained in $\mathcal{C}_{D,\chi}(t + 2\epsilon)$.

The bound on the number of labeled examples now follows directly from known concentration results using the expected number of partitions instead of the maximum in the standard VC-dimension bounds (e.g., see Theorem 21 in Appendix A). This bound ensures that with probability $1-\frac{\delta}{2}$, none of the functions $f \in \mathcal{C}_{D,\chi}(t+2\epsilon)$ with $err(f) \geq \epsilon$ have $\widehat{err}(f) = 0$.

The above two arguments together imply that with probability $1 - \delta$, all $f \in \mathcal{C}$ with $\widehat{err}(f) = 0$ and $\widehat{err}_{unl}(f) \leq t + \epsilon$ have $err(f) \leq \epsilon$, and furthermore c^* has $\widehat{err}_{unl}(c^*) \leq t + \epsilon$. This in turn implies that with probability at least $1 - \delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$ as desired. □

We can also give a bound where we specify the number of labeled examples as a function of the *unlabeled sample*; this is useful because we can imagine our learning algorithm performing some calculations over the unlabeled data and then deciding how many labeled examples to purchase.

THEOREM 9. *If $c^* \in \mathcal{C}$ and $err_{unl}(c^*) = t$, then an unlabeled sample S of size*

$$\mathcal{O}\left(\frac{\max[VCdim(\mathcal{C}), VCdim(\chi(\mathcal{C}))]}{\epsilon^2}\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{2}{\delta}\right)$$

is sufficient so that if we label m_l examples drawn uniformly at random from S, where

$$m_l > \frac{4}{\epsilon}\left[\ln(2\mathcal{C}_{S,\chi}(t+\epsilon)\big[2m_l, \overline{S}\big]) + \ln\frac{4}{\delta}\right]$$

then with probability at least $1-\delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$.

PROOF. Standard VC-bounds (in the same form as for Theorem 8) imply that the number of *labeled* examples m_l is sufficient to guarantee the conclusion of the theorem with "$err(f)$" replaced by "$err_{\overline{S}}(f)$" (the error with respect to $\overline{S}$) and "ϵ" replaced with "$\epsilon/2$". The number of *unlabeled* examples is enough to ensure that, with probability $\geq 1 - \frac{\delta}{2}$, for all $f \in \mathcal{C}$, $|err(f) - err_{\overline{S}}(f)| \leq \epsilon/2$. Combining these two statements yields the theorem. □

Note that if we assume $err_{unl}(c^*) = 0$, then we can use the set $\mathcal{C}_{S,\chi}(0)$ instead of $\mathcal{C}_{S,\chi}(t+\epsilon)$ in the formula giving the number of labeled examples in Theorem 9.

Note: Notice that for the setting of Example 1, in the worst case (over distributions D) this will essentially recover the standard margin sample-complexity bounds for the number of labeled examples. In particular, $\mathcal{C}_{S,\chi}(0)$ contains only those separators that split S with margin $\geq \gamma$, and therefore, $s = \left|\mathcal{C}_{S,\chi}(0)[2m_l, \overline{S}]\right|$ is no greater than the maximum number of ways of splitting $2m_l$ points with margin γ. However, if the distribution is helpful, then the bounds can be much better because there may be many fewer ways of splitting S with margin γ. For instance, in the case of two well-separated "blobs" illustrated in Figure 1, if S is large enough, we would have just $s = 4$.

Theorem 9 immediately implies the following stratified version, which applies to the case in which one repeatedly draws labeled examples until that number is sufficient to justify the most-compatible hypothesis found.

COROLLARY 10. *An unlabeled sample S of size*

$$\mathcal{O}\left(\frac{\max[VCdim(\mathcal{C}), VCdim(\chi(\mathcal{C}))]}{\epsilon^2}\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{2}{\delta}\right)$$

is sufficient so that with probability $\geq 1-\delta$ we have that simultaneously for every $k \geq 0$ the following is true: if we label m_k examples drawn uniformly at random from S, where

$$m_k > \frac{4}{\epsilon}\left[\ln\left(2\mathcal{C}_{S,\chi}((k+1)\epsilon)\big[2m_k, \overline{S}\big]\right) + \ln\frac{4(k+1)(k+2)}{\delta}\right]$$

then all $f \in \mathcal{C}$ with $\widehat{err}(f) = 0$ and $\widehat{err}_{unl}(f) \leq (k+1)\epsilon$ have $err(f) \leq \epsilon$.

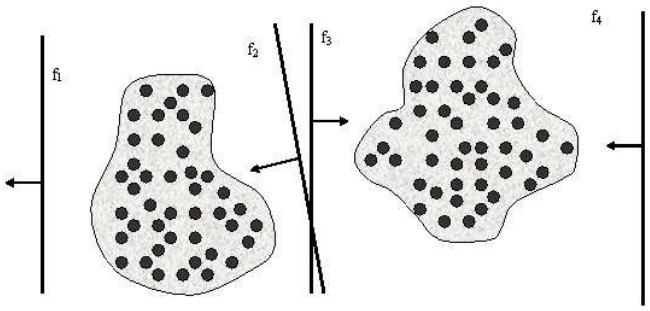

Fig. 1. Linear separators with a margin-based notion of compatibility. If the distribution is uniform over two well-separated "blobs" and the unlabeled set S is sufficiently large, the set $\mathcal{C}_{S,\chi}(0)$ contains only four different partitions of S, shown in the figure as f_1, f_2, f_3, and f_4. Therefore, Theorem 9 implies that we only need $O(1/\epsilon)$ labeled examples to learn well.

Interpretation: This corollary is an analog of Theorem 6 and it justifies a stratification based on the estimated unlabeled error rates. That is, beginning with $k = 0$, one draws the specified number of examples and checks to see if a sufficiently compatible hypothesis can be found. If so, one halts with success, and if not, one increments k and tries again. Since $k \leq \frac{1}{\epsilon}$, we clearly have a fallback property: the number of labeled examples required is never much worse than the number of labeled examples required by a standard supervised learning algorithm.

If one does not have the ability to draw additional labeled examples, then we can fix m_l and instead stratify over estimation error as in [Bartlett et al. 1999]. We discuss this further in our agnostic bounds in Section 3.1.3 below.

3.1.3 *The agnostic case.* The bounds given so far have been based on the assumption that the target function belongs to $\mathcal{C}$ (so that we can assume there will exist $f \in \mathcal{C}$ with $\widehat{err}(f) = 0$). One can also derive analogous results for the agnostic (unrealizable) case, where we do not make that assumption. We first present one immediate bound of this form, and then show how we can use it in order to trade off labeled and unlabeled error in a near-optimal way. We also discuss the relation of this to a common "regularization" technique used in semi-supervised learning. As we will see, the differences between these two point to certain potential pitfalls in the standard regularization approach.

THEOREM 11. *Let* $f_t^* = \text{argmin}_{f \in \mathcal{C}}[err(f) | err_{unl}(f) \leq t]$. *Then an unlabeled sample* S *of size*

$$\mathcal{O}\left(\frac{\max[VCdim(\mathcal{C}), VCdim(\chi(\mathcal{C}))]}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{2}{\delta}\right)$$

and a labeled sample of size

$$m_l \geq \frac{8}{\epsilon^2}\left[\log\left(2\mathcal{C}_{D,\chi}(t+2\epsilon)[2m_l, D]\right) + \log\frac{4}{\delta}\right]$$

is sufficient so that with probability $\geq 1-\delta$, *the* $f \in \mathcal{C}$ *that optimizes* $\widehat{err}(f)$ *subject to* $\widehat{err}_{unl}(f) \leq t + \epsilon$ *has* $err(f) \leq err(f_t^*) + \epsilon + \sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*) + 2\epsilon$.

PROOF. The given unlabeled sample size implies that with probability $1-\delta/2$, all $f \in \mathcal{C}$ have $|\widehat{err}_{unl}(f) - err_{unl}(f)| \leq \epsilon$, which also implies that $\widehat{err}_{unl}(f_t^*) \leq t+\epsilon$. The labeled sample size, using standard VC bounds (e.g, Theorem 22 in the Appendix B) imply that with probability at least $1-\delta/4$, all $f \in \mathcal{C}_{D,\chi}(t+2\epsilon)$ have $|\widehat{err}(f) - err(f)| \leq \epsilon$. Finally, by Hoeffding bounds, with probability at least $1-\delta/4$ we have

$$\widehat{err}(f_t^*) \leq err(f_t^*) + \sqrt{\log(4/\delta)/(2m_l)}.$$

Therefore, with probability at least $1-\delta$, the $f \in \mathcal{C}$ that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(f) \leq t+\epsilon$ has

$$err(f) \leq \widehat{err}(f)+\epsilon \leq \widehat{err}(f_t^*)+\epsilon \leq err(f_t^*)+\epsilon+\sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*)+2\epsilon,$$

as desired. □

Interpretation: Given a value t, Theorem 11 bounds the number of labeled examples needed to achieve error at most ϵ larger than that of the best function f_t^* of unlabeled error rate at most t. Alternatively, one can also state Theorem 11 in the form more commonly used in statistical learning theory: given *any* number of labeled examples m_l and given $t > 0$, Theorem 11 implies that with high probability, the function f that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(f) \leq t+\epsilon$ satisfies

$$err(f) \leq \widehat{err}(f) + \epsilon_t \leq err(f_t^*) + \epsilon_t + \sqrt{\frac{\log(4/\delta)}{2m_l}}$$

where

$$\epsilon_t = \sqrt{\frac{8}{m_l}\log\left(8\mathcal{C}_{D,\chi}(t+2\epsilon)[2m_l, D]/\delta\right)}.$$

Note that as usual, there is an inherent tradeoff here between the quality of the comparison function f_t^*, which improves as t increases, and the estimation error ϵ_t, which gets worse as t increases. Ideally, one would like to achieve a bound of $\min_t[err(f_t^*)+\epsilon_t] + \sqrt{\log(4/\delta)/(2m_l)}$; i.e., as if the optimal value of t were known in advance. We can perform nearly as well as this bound by (1) performing a stratification over t (so that the bound holds simultaneously for all values of t) and (2) using an estimate $\hat{\epsilon}_t$ of ϵ_t that we can calculate from the unlabeled sample and therefore use in the optimization. In particular, letting $f_t = \text{argmin}_{f' \in \mathcal{C}}[\widehat{err}(f') : \widehat{err}_{unl}(f') \leq t]$, we will output $f = \text{argmin}_{f_t}[\widehat{err}(f_t) + \hat{\epsilon}_t]$.

Specifically, given a set S of unlabeled examples and m_l labeled examples, let $\hat{\epsilon}_t = \hat{\epsilon}_t(S, m_l) = \sqrt{\frac{24}{m_l}\log\left(8\mathcal{C}_{S,\chi}(t)[m_l, S]\right)}$, where we define $\mathcal{C}_{S,\chi}(t)[m_l, S]$ to be the number of different partitions of the first m_l points in S using functions in $\mathcal{C}_{S,\chi}(t)$, i.e., using functions of empirical unlabeled error at most t (we assume $|S| \geq m_l$). Then we have the following theorem.

THEOREM 12. *Let $f_t^* = \text{argmin}_{f' \in \mathcal{C}}[err(f') | err_{unl}(f') \leq t]$ and define $\hat{\epsilon}(f') = \hat{\epsilon}_{t'}$ for $t' = \widehat{err}_{unl}(f')$. Then, given m_l labeled examples, with probability at least $1-\delta$, the function*

$$f = \text{argmin}_{f'}[\widehat{err}(f') + \hat{\epsilon}(f')]$$

satisfies the guarantee that

$$err(f) \leq \min_t [err(f_t^*) + \hat{\epsilon}(f_t^*)] + 5\sqrt{\frac{\log(8/\delta)}{m_l}}$$

PROOF. First we argue that with probability at least $1-\delta/2$, for all $f' \in \mathcal{C}$ we have $err(f') \leq \widehat{err}(f') + \hat{\epsilon}(f') + 4\sqrt{\frac{\log(8/\delta)}{m_l}}$. In particular, define $\mathcal{C}_0 = \mathcal{C}_{S,\chi}(0)$ and inductively for $k > 0$ define $\mathcal{C}_k = \mathcal{C}_{S,\chi}(t_k)$ for t_k such that $\mathcal{C}_k[m_l, S] = 8\mathcal{C}_{k-1}[m_l, S]$. (If necessary, arbitrarily order the functions with empirical unlabeled error exactly t_k and choose a prefix such that the size condition holds.) Also, we may assume without loss of generality that $\mathcal{C}_0[m_l, S] \geq 1$. Then, using bounds of [Boucheron et al. 2000] (see also Appendix A), we have that with probability at least $1-\delta/2^{k+2}$, all $f' \in C_k \setminus C_{k-1}$ satisfy:

$$\begin{aligned}
err(f') &\leq \widehat{err}(f') + \sqrt{\frac{6}{m_l}\log(\mathcal{C}_k[m_l, S])} + 4\sqrt{\frac{1}{m_l}\log(2^{k+3}/\delta)} \\
&\leq \widehat{err}(f') + \sqrt{\frac{6}{m_l}\log(\mathcal{C}_k[m_l, S])} + 4\sqrt{\frac{1}{m_l}\log(2^k)} + 4\sqrt{\frac{1}{m_l}\log(8/\delta)} \\
&\leq \widehat{err}(f') + \sqrt{\frac{6}{m_l}\log(\mathcal{C}_k[m_l, S])} + \sqrt{\frac{6}{m_l}\log(8^k)} + 4\sqrt{\frac{1}{m_l}\log(8/\delta)} \\
&\leq \widehat{err}(f') + 2\sqrt{\frac{6}{m_l}\log(\mathcal{C}_k[m_l, S])} + 4\sqrt{\frac{1}{m_l}\log(8/\delta)} \\
&\leq \widehat{err}(f') + \hat{\epsilon}(f') + 4\sqrt{\frac{1}{m_l}\log(8/\delta)}.
\end{aligned}$$

Now, let $f^* = \text{argmin}_{f_t^*}[err(f_t^*) + \hat{\epsilon}(f_t^*)]$. By Hoeffding bounds, with probability at least $1-\delta/2$ we have $\widehat{err}(f^*) \leq err(f^*) + \sqrt{\log(2/\delta)/(2m_l)}$. Also, by construction we have $\widehat{err}(f) + \hat{\epsilon}(f) \leq \widehat{err}(f^*) + \hat{\epsilon}(f^*)$. Therefore with probability at least $1-\delta$ we have:

$$\begin{aligned}
err(f) &\leq \widehat{err}(f) + \hat{\epsilon}(f) + 4\sqrt{\log(8/\delta)/m_l} \\
&\leq \widehat{err}(f^*) + \hat{\epsilon}(f^*) + 4\sqrt{\log(8/\delta)/m_l} \\
&\leq err(f^*) + \hat{\epsilon}(f^*) + 5\sqrt{\log(8/\delta)/m_l}
\end{aligned}$$

as desired. □

The above result bounds the error of the function f produced in terms of the quantity $\hat{\epsilon}(f^*)$ which depends on the *empirical* unlabeled error rate of f^*. If our unlabeled sample S is sufficiently large to estimate all unlabeled error rates to $\pm\epsilon$, then with high probability we have $\widehat{err}(f_t^*) \leq t + \epsilon$, so $\hat{\epsilon}(f_t^*) \leq \hat{\epsilon}_{t+\epsilon}$, and moreover $\mathcal{C}_{S,\chi}(t+\epsilon) \subseteq \mathcal{C}_{D,\chi}(t+2\epsilon)$. So, our error term $\hat{\epsilon}(f_t^*)$ is at most $\sqrt{\frac{24}{m_l}\log\left(8\mathcal{C}_{D,\chi}(t+2\epsilon)[m_l, S]\right)}$. Recall that our ideal error term ϵ_t for the case that t was given to the algorithm in advance, factoring out the dependence on δ, was $\sqrt{\frac{8}{m_l}\log\left(8\mathcal{C}_{D,\chi}(t+2\epsilon)[2m_l, D]\right)}$. [Boucheron et al. 2000] show that for any class $\mathcal{C}$, the quantity $\log(\mathcal{C}[m, S])$ is tightly concentrated about $\log(\mathcal{C}[m, D])$ (see

also Theorem 25 in the Appendix B), so up to multiplicative constants, these two bounds are quite close.

Interpretation and use of unlabeled error rate as a regularizer: The above theorem suggests to optimize the sum of the empirical labeled error rate and an estimation-error bound based on the unlabeled error rate. A common related approach used in practice in machine learning (e.g., [Chapelle et al. 2006]) is to just directly optimize the sum of the two kinds of error: i.e., to find $\text{argmin}_f[\widehat{err}(f) + \widehat{err}_{unl}(f)]$. However, this is not generically justified in our framework, because the labeled and unlabeled error rates are really of different "types". In particular, depending on the concept class and notion of compatibility, a small change in unlabeled error rate could substantially change the size of the compatible set.[3] For example, suppose all functions in $\mathcal{C}$ have unlabeled error rate 0.6, except for two: function f_0 has unlabeled error rate 0 and labeled error rate 1/2, and function $f_{0.5}$ has unlabeled error rate 0.5 and labeled error rate 1/10. Suppose also that $\mathcal{C}$ is sufficiently large that with high probability it contains some functions f that drastically overfit, giving $\widehat{err}(f) = 0$ even though their true error is close to 1/2. In this case, we would like our algorithm to pick out $f_{0.5}$ (since its labeled error rate is fairly low, and we cannot trust the functions of unlabeled error 0.6). However, even if we use a regularization parameter λ, there is no way to make $f_{0.5} = \text{argmin}_f[\widehat{err}(f) + \lambda err_{unl}(f)]$: in particular, one cannot have $1/10 + 0.5\lambda \leq \min[1/2 + 0\lambda, 0 + 0.6\lambda]$. So, in this case, this approach will not have the desired behavior.

Note: One could further derive tighter bounds, both in terms of labeled and unlabeled examples, that are based on other distribution dependent complexity measures and using stronger concentration results (see e.g. [Boucheron et al. 2005]).

3.2 ϵ-Cover-based Bounds

The results in the previous section are uniform convergence bounds: they provide guarantees for *any* algorithm that optimizes over the observed data. In this section, we consider stronger bounds based on ϵ-covers that apply to algorithms that behave in a specific way: they first use the unlabeled examples to choose a "representative" set of compatible hypotheses, and then use the labeled sample to choose among these. Bounds based on ϵ-covers exist in the classical PAC setting, but in our framework these bounds and algorithms of this type are especially natural, and the bounds are often much lower than what can be achieved via uniform convergence. For simplicity, we restrict ourselves in this section to the realizable case. However one can combine ideas in Section 3.1.3 with ideas in this section in order to derive bounds in the agnostic case as well. We first present our generic bounds. In Section 3.2.1 we discuss natural settings in which they can be especially useful, and in then Section 3.2.2 we present even tighter bounds for co-training.

Recall that a set $C_\epsilon \subseteq 2^{\mathcal{X}}$ is an ϵ-cover for $\mathcal{C}$ with respect to D if for every $f \in \mathcal{C}$ there is a $f' \in C_\epsilon$ which is ϵ-close to f. That is, $\mathbf{Pr}_{x\sim D}(f(x) \neq f'(x)) \leq \epsilon$.

We start with a theorem that relies on knowing a good upper bound on the unlabeled error rate of the target function $err_{unl}(c^*)$.

[3] On the other hand, for certain compatibility notions and under certain natural assumptions, one can use unlabeled error rate directly, e.g., see e.g., [Sridharan and Kakade 2008].

THEOREM 13. *Assume $c^* \in \mathcal{C}$ and let p be the size of a minimum ϵ-cover for $\mathcal{C}_{D,\chi}(err_{unl}(c^*) + 2\epsilon)$. Then using m_u unlabeled examples and m_l labeled examples for*

$$m_u = \mathcal{O}\left(\frac{\max[VCdim(\mathcal{C}), VCdim(\chi(\mathcal{C}))]}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{2}{\delta}\right) \text{ and } m_l = \mathcal{O}\left(\frac{1}{\epsilon}\ln\frac{p}{\delta}\right),$$

we can with probability $1-\delta$ identify a hypothesis $f \in \mathcal{C}$ with $err(f) \leq 6\epsilon$.

PROOF. Let $t = err_{unl}(c^*)$. Now, given the unlabeled sample S_U, define $\mathcal{C}' \subseteq \mathcal{C}$ as follows: for every labeling of S_U that is consistent with some f in $\mathcal{C}$, choose a hypothesis in $\mathcal{C}$ for which $\widehat{err}_{unl}(f)$ is smallest among all the hypotheses corresponding to that labeling. Next, we obtain C_ϵ by eliminating from $\mathcal{C}'$ those hypotheses f with the property that $\widehat{err}_{unl}(f) > t + \epsilon$. We then apply a greedy procedure on C_ϵ to obtain $G_\epsilon = \{g_1, \cdots, g_s\}$, as follows:

Initialize $C_\epsilon^1 = C_\epsilon$ and $i = 1$.

(1) Let $g_i = \operatorname{argmin}_{f \in C_\epsilon^i} \widehat{err}_{unl}(f)$.

(2) Using the unlabeled sample S_U, determine C_ϵ^{i+1} by deleting from C_ϵ^i those hypotheses f with the property that $\hat{d}(g_i, f) < 3\epsilon$.

(3) If $C_\epsilon^{i+1} = \emptyset$ then set $s = i$ and stop; else, increase i by 1 and goto 1.

We now show that with high probability, G_ϵ is a 5ϵ-cover of $\mathcal{C}_{D,\chi}(t)$ with respect to D and has size at most p. First, our bound on m_u is sufficient to ensure that with probability $\geq 1 - \frac{\delta}{2}$, we have (a) $|\hat{d}(f,g) - d(f,g)| \leq \epsilon$ for all $f, g \in \mathcal{C}$ and (b) $|\widehat{err}_{unl}(f) - err_{unl}(f)| \leq \epsilon$ for all $f \in \mathcal{C}$. Let us assume in the remainder that this (a) and (b) are indeed satisfied. Now, (a) implies that any two functions in $\mathcal{C}$ that agree on S_U have distance at most ϵ, and therefore $\mathcal{C}'$ is an ϵ-cover of $\mathcal{C}$. Using (b), this in turn implies that C_ϵ is an ϵ-cover for $\mathcal{C}_{D,\chi}(t)$. By construction, G_ϵ is a 3ϵ-cover of C_ϵ with respect to distribution $\overline{S_U}$, and thus (using (a)) G_ϵ is a 4ϵ-cover of C_ϵ with respect to D, which implies that G_ϵ is a 5ϵ-cover of $\mathcal{C}_{D,\chi}(t)$ with respect to D.

We now argue that G_ϵ has size at most p. Fix some optimal ϵ-cover $\{f_1, \ldots, f_p\}$ of $\mathcal{C}_{D,\chi}(err_{unl}(c^*) + 2\epsilon)$. Consider function g_i and suppose that g_i is covered by $f_{\sigma(i)}$. Then the set of functions deleted in step (2) of the procedure include those functions f satisfying $d(g_i, f) < 2\epsilon$ which by triangle inequality includes those satisfying $d(f_{\sigma(i)}, f) \leq \epsilon$. Therefore, the set of functions deleted include those covered by $f_{\sigma(i)}$ and so for all $j > i$, $\sigma(j) \neq \sigma(i)$; in particular, σ is 1-1. This implies that G_ϵ has size at most p.

Finally, to learn c^* we simply output the function $f \in G_\epsilon$ of lowest empirical error over the labeled sample. By Chernoff bounds, the number of labeled examples is enough to ensure that with probability $\geq 1 - \frac{\delta}{2}$ the empirical optimum hypothesis in G_ϵ has true error at most 6ϵ. This implies that overall, with probability $\geq 1 - \delta$, we find a hypothesis of error at most 6ϵ. □

Note that Theorem 13 relies on knowing a good upper bound on $err_{unl}(c^*)$. If we do not have such an upper bound, then one can perform a stratification as in Sections 3.1.2 and 3.1.3. For example, if we have a desired maximum error rate ϵ and we do not know a good upper bound for $err_{unl}(c^*)$ but we have the

ability to draw additional labeled examples as needed, then we can simply run the procedure in Theorem 13 for various value of p, testing on each round to see if the hypothesis f found indeed has zero empirical labeled error rate. One can show that $m_l = \mathcal{O}\left(\frac{1}{\epsilon}\ln\frac{p}{\delta}\right)$ labeled examples are sufficient in total for all the "validation" steps.[4] If the number of labeled examples m_l is fixed, then one can also perform a stratification over the target error ϵ.

3.2.1 *Some illustrative examples.* To illustrate the power of ϵ-cover bounds, we now present two examples where these bounds allow for learning from significantly fewer labeled examples than is possible using uniform convergence.

Graph-based learning: Consider the setting of graph-based algorithms (e.g., Example 3). In particular, the input is a graph G where each node is an example and $\mathcal{C}$ is the class of all boolean functions over the nodes of G. Let us define the incompatibility of a hypothesis to be the fraction of edges in G cut by it. Suppose now that the graph G consists of two cliques of $n/2$ vertices, connected together by $\epsilon n^2/4$ edges. Suppose the target function c^* labels one of the cliques as positive and one as negative, so the target function indeed has unlabeled error rate less than ϵ. Now, given any set S_L of $m_l < \epsilon n/4$ labeled examples, there is always a highly-compatible hypothesis consistent with S_L that just separates the positive points in S_L from the entire rest of the graph: the number of edges cut will be at most $nm_l < \epsilon n^2/4$. However, such a hypothesis has true error nearly 1/2 since it has less than $\epsilon n/4$ positive examples. So, we do not yet have uniform convergence over the space of highly compatible hypotheses, since this hypothesis has zero empirical error but high true error. Indeed, this illustrates an overfitting problem that can occur with a direct minimum-cut approach to learning [Blum and Chawla 2001; Joachims 2003; Blum et al. 2004]. On the other hand, the set of functions of unlabeled error rate less than ϵ has a small ϵ-cover: in particular, *any* partition of G that cuts less than $\epsilon n^2/4$ edges must be ϵ-close to (a) the all-positive function, (b) the all-negative function, (c) the target function c^*, or (d) the complement of the target function $1 - c^*$. So, ϵ-cover bounds act as if the concept class had only 4 functions and so by Theorem 13 we need only $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$ labeled examples to learn well.[5] (In fact, since the functions in the cover are all far from each other, we really need only $O(\log\frac{1}{\delta})$ examples. This issue is explored further in Theorem 15).

Simple co-training: For another case where ϵ-cover bounds can beat uniform-convergence bounds, imagine examples are *pairs* of points in $\{0,1\}^d$, $\mathcal{C}$ is the class of linear separators, and compatibility is determined by whether both points are on the same side of the separator (i.e., the case of Example 4). Now suppose for simplicity that the target function just splits the hypercube on the first coordinate, and the distribution is uniform over pairs having the same first coordinate (so the

[4]Specifically, note that as we increase t (our current estimate for the unlabeled error rate of the target function), the associated p (which is an integer) increases in discrete jumps, $p_1, p_2, \ldots$. We can then simply spread the "δ" parameter across the different runs, in particular run i would use $\delta/i(i+1)$. Since $p_i \geq i$, this implies that $m_l = \mathcal{O}\left(\frac{1}{\epsilon}\ln\frac{p}{\delta}\right)$ labeled examples are sufficient for all the "validation" steps.

[5]Effectively, ϵ-cover bounds allow one to rule out a hypothesis that, say, just separates the positive points in S_L from the rest of the graph by noting that this hypothesis is very close (with respect to D) to the all-negative hypothesis, and *that* hypothesis has a high labeled-error rate.

target is fully compatible). We then have the following.

THEOREM 14. *Given poly(d) unlabeled examples S_U and $\frac{1}{4}\log d$ labeled examples S_L, with high probability there will exist functions of true error $1/2 - 2^{-\frac{1}{2}\sqrt{d}}$ that are consistent with S_L and compatible with S_U.*

PROOF. Let V be the set of all variables (not including x_1) that (a) appear in *every* positive example of S_L and (b) appear in *no* negative example of S_L. In other words, these are variables x_i such that the function $f(x) = x_i$ correctly classifies all examples in S_L. Over the draw of S_L, each variable has a $(1/2)^{2|S_L|} = 1/\sqrt{d}$ chance of belonging to V, so the expected size of V is $(d-1)/\sqrt{d}$ and so by Chernoff bounds, with high probability V has size at least $\frac{1}{2}\sqrt{d}$. Now, consider the hypothesis corresponding to the conjunction of all variables in V. This correctly classifies the examples in S_L, and with probability at least $1-2|S_U|2^{-|V|}$ it classifies *every* other example in S_U negative because each example in S_U has only a $1/2^{|V|}$ chance of satisfying every variable in V. Since $|S_U| = \text{poly}(d)$, this means that with high probability this conjunction is compatible with S_U and consistent with S_L, even though its true error is at least $1/2 - 2^{-\frac{1}{2}\sqrt{d}}$. □

So, given only a set S_U of poly(d) unlabeled examples and a set S_L of $\frac{1}{4}\log d$ labeled examples we would not want to use a uniform convergence based algorithm since we do not yet have uniform convergence. In contrast, the cover-size of the set of functions compatible with S_U is constant, so ϵ-cover based bounds again allow learning from just only $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$ labeled examples (Theorem 13). In fact as we show in Theorem 15 we only need $\mathcal{O}\left(\log_{\frac{1}{\epsilon}}\frac{1}{\delta}\right)$ labeled examples in this case.

3.2.2 *Learning from even fewer labeled examples.* In some cases, unlabeled data can allow us to learn from even fewer labeled examples than given by Theorem 13. In particular, consider a co-training setting where the target c^* is fully compatible *and* D satisfies the property that the two views x_1 and x_2 are conditionally independent given the label $c^*(\langle x_1, x_2\rangle)$. As shown by [Blum and Mitchell 1998], one can boost any weak hypothesis from unlabeled data in this setting (assuming one has enough labeled data to produce a weak hypothesis). Related sample complexity results are given in [Dasgupta et al. 2001]. In fact, we can use the notion of ϵ-covers to show that we can learn from just a single labeled example. Specifically, for any concept classes $\mathcal{C}_1$ and $\mathcal{C}_2$, we have:

THEOREM 15. *Assume that $err(c^*) = err_{unl}(c^*) = 0$ and D satisfies independence given the label. Then for any $\tau \le \epsilon/4$, using m_u unlabeled examples and m_l labeled examples we can find a hypothesis that with probability $1-\delta$ has error at most ϵ, for*

$$m_u = \mathcal{O}\left(\frac{1}{\tau}\left[(VCdim(\mathcal{C}_1) + VCdim(\mathcal{C}_2))\ln\frac{1}{\tau} + \ln\frac{2}{\delta}\right]\right) \quad \text{and} \quad m_l = \mathcal{O}\left(\log_{\frac{1}{\tau}}\frac{1}{\delta}\right).$$

PROOF. We will assume for simplicity the setting of Example 3, where $c^* = c_1^* = c_2^*$ and also $D_1 = D_2 = \tilde{D}$ (the general case is handled similarly, but just requires more notation).

We start by characterizing the hypotheses with low unlabeled error rate. Recall that $\chi(f, D) = \mathbf{Pr}_{\langle x_1,x_2\rangle\sim D}[f(x_1) = f(x_2)]$, and for concreteness assume f predicts

using x_1 if $f(x_1) \neq f(x_2)$. Consider $f \in \mathcal{C}$ with $err_{unl}(f) \leq \tau$ and let's define $p_- = \mathbf{Pr}_{x \in \tilde{D}}[c^*(x) = 0]$, $p_+ = \mathbf{Pr}_{x \in \tilde{D}}[c^*(x) = 1]$ and for $i, j \in \{0, 1\}$ define $p_{ij} = \mathbf{Pr}_{x \in \tilde{D}}[f(x) = i, c^*(x) = j]$. We clearly have $err(f) = p_{10} + p_{01}$. From $err_{unl}(f) = \mathbf{Pr}_{(x_1,x_2)\sim D}[f(x_1) \neq f(x_2)] \leq \tau$, using the independence given the label of D, we get

$$\frac{2p_{10}p_{00}}{p_{10} + p_{00}} + \frac{2p_{01}p_{11}}{p_{01} + p_{11}} \leq \tau.$$

In particular, the fact that $\frac{2p_{10}p_{00}}{p_{10}+p_{00}} \leq \tau$ implies that we cannot have both $p_{10} > \tau$ and $p_{00} > \tau$, and the fact that $\frac{2p_{01}p_{11}}{p_{01}+p_{11}} \leq \tau$ implies that we cannot have both $p_{01} > \tau$ and $p_{11} > \tau$. Therefore, any hypothesis f with $err_{unl}(f) \leq \tau$ falls in one of the following categories:

(1) f is "close to c^*": $p_{10} \leq \tau$ and $p_{01} \leq \tau$; so $err(f) \leq 2\tau$.
(2) f is "close to $\overline{c^*}$": $p_{00} \leq \tau$ and $p_{11} \leq \tau$; so $err(f) \geq 1 - 2\tau$.
(3) f "almost always predicts negative": for $p_{10} \leq \tau$ and $p_{11} \leq \tau$; so $\mathbf{Pr}[f(x) = 0] \geq 1 - 2\tau$.
(4) f "almost always predicts positive": for $p_{00} \leq \tau$ and $p_{01} \leq \tau$; so $\mathbf{Pr}[f(x) = 0] \leq 2\tau$.

Let f_1 be the constant positive function and f_0 be the constant negative function. Now note that our bound on m_u is sufficient to ensure that with probability $\geq 1 - \frac{\delta}{2}$, we have (a) $|\hat{d}(f, g) - d(f, g)| \leq \tau$ for all $f, g \in \mathcal{C}$ and (b) all $f \in \mathcal{C}$ with $\widehat{err}_{unl}(f) = 0$ satisfy $err_{unl}(f) \leq \tau$. Let us assume in the remainder that this (a) and (b) are indeed satisfied. By our previous analysis, there are at most four kinds of hypotheses consistent with unlabeled data: those close to c^*, those close to its complement $\overline{c^*}$, those close to f_0, and those close to f_1. Furthermore, c^*, $\overline{c^*}$, f_0, and f_1 *are* compatible with the unlabeled data.

So, algorithmically, we first check to see if there exists a hypothesis $g \in \mathcal{C}$ with $\widehat{err}_{unl}(g) = 0$ such that $\hat{d}(f_1, g) \geq 3\tau$ and $\hat{d}(f_0, g) \geq 3\tau$. If such a hypothesis g exists, then it must satisfy either case (1) or (2) above. Therefore, we know that one of $\{g, \overline{g}\}$ is 2τ-close to c^*. If not, we must have $p_+ \leq 4\tau$ or $p_- \leq 4\tau$, in which case we know that one of $\{f_0, f_1\}$ is 4τ-close to c^*. So, either way we have a set of two functions, opposite to each other, one of which is at least 4τ-close to c^*. We finally use $O(\log_{\frac{1}{\tau}} \frac{1}{\delta})$ labeled examples to pick one of these to output, namely the one with lowest empirical labeled error. Lemma 16 below then implies that with probability $1 - \delta$ the function we output has error at most $4\tau \leq \epsilon$. □

LEMMA 16. *Consider $\tau < \frac{1}{8}$. Let $C_\tau = \{f, \overline{f}\}$ be a subset of $\mathcal{C}$ containing two opposite hypotheses with the property that one of them is τ-close to c^*. Then, $m_l > 6 \log_{(\frac{1}{\tau})}(\frac{1}{\delta})$ labeled examples are sufficient so that with probability $\geq 1 - \delta$, the concept in C_τ that is τ-close to c^* in fact has lower empirical error.*

PROOF. See Appendix B. □

In particular, by reducing τ to poly(δ) in Theorem 15, we can reduce the number of labeled examples needed m_l to *one*. Note however that we will need polynomially more unlabeled examples.

In fact, the result in Theorem 15 can be extended to the case that D^+ and D^- merely satisfy constant expansion rather than full independence given the label, see [Balcan et al. 2004].

Note: Theorem 15 illustrates that if data is especially well behaved with respect to the compatibility notion, then our bounds on labeled data can be extremely good. In Section 4.2, we show for the case of linear separators and independence given the label, we can give *efficient* algorithms, achieving the bounds in Theorem 15 in terms of labeled examples by a polynomial time algorithm. Note, however, that both these bounds rely heavily on the assumption that the target is fully compatible. If the assumption is more of a "hope" than a belief, then one would need an additional sample of $1/\epsilon$ labeled examples just to validate the hypothesis produced.

4. ALGORITHMIC RESULTS

In this section we give several examples of *efficient* algorithms in our model that are able to learn using sample sizes comparable to those described in Section 3. Note that our focus is on achieving a low-error hypothesis (also called minimizing 0-1 loss). Another common practice in machine learning (both in the context of supervised and semi-supervised learning) is to instead try to minimize a surrogate convex loss that is easier to optimize [Chapelle et al. 2006]. While this does simplify the computational problem, it does not in general solve the true goal of achieving low error.

4.1 A simple case

We give here a simple example to illustrate the bounds in Section 3.1.1, and for which we can give a polynomial-time algorithm that takes advantage of them. Let the instance space $\mathcal{X} = \{0,1\}^d$, and for $x \in \mathcal{X}$, let $\mathsf{vars}(x)$ be the set of variables set to 1 by x. Let $\mathcal{C}$ be the class of monotone disjunctions (e.g., $x_1 \vee x_3 \vee x_6$), and for $f \in \mathcal{C}$, let $\mathsf{vars}(f)$ be the set of variables disjoined by f. Now, suppose we say an example x is compatible with function f if either $\mathsf{vars}(x) \subseteq \mathsf{vars}(f)$ or else $\mathsf{vars}(x) \cap \mathsf{vars}(f) = \phi$. This is a very strong notion of "margin": it says, in essence, that every variable is either a positive indicator or a negative indicator, and no example should contain both positive and negative indicators.

Given this setup, we can give a simple $\mathrm{PAC}_{\mathrm{SSL}}$-learning algorithm for this pair $(\mathcal{C}, \chi)$: that is, an algorithm with sample size bounds that are polynomial (or in this case, matching) those in Theorem 4. Specifically, we can prove the following:

THEOREM 17. *The class $\mathcal{C}$ of monotone disjunctions is* $\mathrm{PAC}_{\mathrm{SSL}}$*-learnable under the compatibility notion defined above.*

PROOF. We begin by using our unlabeled data to construct a graph on d vertices (one per variable), putting an edge between two vertices i and j if there is any example x in our unlabeled sample with $i, j \in \mathsf{vars}(x)$. We now use our labeled data to label the components. If the target function is fully compatible, then no component will get multiple labels (if some component does get multiple labels, we halt with failure). Finally, we produce the hypothesis f such that $\mathsf{vars}(f)$ is the union of the positively-labeled components. This is fully compatible with the unlabeled data and has zero error on the labeled data, so by Theorem 4, if the sizes

of the data sets are as given in the bounds, with high probability the hypothesis produced will have error at most ϵ. □

Notice that if we want to view the algorithm as "purchasing" labeled data, then we can simply examine the graph, count the number of connected components k, and then request $\frac{1}{\epsilon}[k \ln 2 + \ln \frac{2}{\delta}]$ labeled examples. (Here, $2^k = |\mathcal{C}_{S,\chi}(0)|$.) By the proof of Theorem 4, with high probability $2^k \leq |\mathcal{C}_{D,\chi}(\epsilon)|$, so we are purchasing no more than the number of labeled examples in the theorem statement.

Also, it is interesting to see the difference between a "helpful" and "non-helpful" distribution for this problem. An especially *non*-helpful distribution would be the uniform distribution over all examples x with $|\mathsf{vars}(x)| = 1$, in which there are d components. In this case, unlabeled data does not help at all, and one still needs $\Omega(d)$ labeled examples (or, even $\Omega\left(\frac{d}{\epsilon}\right)$ if the distribution is non-uniform as in the lower bounds of [Ehrenfeucht et al. 1989]). On the other hand, a helpful distribution is one such that with high probability the number of components is small, such as the case of features appearing independently given the label.

4.2 Co-training with linear separators

We now consider the case of co-training where the hypothesis class is the class of linear separators. For simplicity we focus first on the case of Example 4: the target function is a linear separator in R^d and each example is a *pair* of points, both of which are assumed to be on the same side of the separator (i.e., an example is a line-segment that does not cross the target hyperplane). We then show how our results can be extended to the more general setting.

As in the previous example, a natural approach is to try to solve the "consistency" problem: given a set of labeled and unlabeled data, our goal is to find a separator that is consistent with the labeled examples and compatible with the unlabeled ones (i.e., it gets the labeled data correct and doesn't cut too many edges). Unfortunately, this consistency problem is NP-hard: given a graph G embedded in R^d with two distinguished points s and t, it is NP-hard to find the linear separator with s on one side and t on the other that cuts the minimum number of edges, *even if the minimum is zero* [Flaxman 2003]. For this reason, we will make an additional assumption, that the two points in an example are each drawn *independently given the label*. That is, there is a single distribution $\tilde{D}$ over R^d, and with some probability p_+, two points are drawn i.i.d. from $\tilde{D}_+$ ($\tilde{D}$ restricted to the positive side of the target function) and with probability $1 - p_+$, the two are drawn i.i.d from $\tilde{D}_-$ ($\tilde{D}$ restricted to the negative side of the target function). Note that our sample complexity results in Section 3.2 extend to weaker assumptions such as distributional expansion introduced by [Balcan et al. 2004], but we need true independence for our algorithmic results. [Blum and Mitchell 1998] also give positive algorithmic results for co-training when (a) the two views of an example are drawn independently given the label (which we are assuming now), (b) the underlying function is learnable via Statistical Query algorithms[6] (which is true for linear separators [Blum et al. 1998]), and (c) we have enough labeled data to produce a weakly-useful

[6]For a detailed description of the Statistical Query model see [Kearns 1998] and [Kearns and Vazirani 1994].

hypothesis (defined below) on one of the views to begin with. We give here an improvement over that result by showing how we can run the algorithm in [Blum and Mitchell 1998] with only *a single* labeled example, thus obtaining an efficient algorithm in our model. It is worth noticing that in the process, we also somewhat simplify the results of [Blum et al. 1998] on efficiently learning linear separators with noise without a margin assumption.

For the analysis below, we need the following definition. A *weakly-useful* predictor is a function f such that for some α that is at least inverse polynomial in the input size we have:

$$\mathbf{Pr}[f(x) = 1|c^*(x) = 1] > \mathbf{Pr}[f(x) = 1|c^*(x) = 0] + \alpha.$$

It is equivalent to the usual notion of a "weak hypothesis" [Kearns and Vazirani 1994] when the target function is balanced, but requires the hypothesis give more information when the target function is unbalanced [Blum and Mitchell 1998]. Also, we will assume for convenience that the target separator passes through the origin, and let us denote the separator by $c^* \cdot x = 0$.

We now describe an efficient algorithm to learn to any desired error rate ϵ in this setting from just a single labeled example. For clarity, we first describe an algorithm whose running time depends polynomially on both the dimension d and $1/\gamma$, where γ is a soft *margin* of separation between positive and negative examples. Formally, in this case we assume that at least some non-negligible probability mass of examples x satisfy $\frac{|x \cdot c^*|}{|x||c^*|} \geq \gamma$; i.e., they have distance at least γ to the separating hyperplane $x \cdot c^* = 0$ after normalization. This is a common type of assumption in machine learning (in fact, often one makes the much stronger assumption that *nearly all* probability mass is on examples x satisfying this condition). We then show how one can replace the dependence on $1/\gamma$ with instead a polynomial dependence on the number of bits of precision b in the data, using the Outlier Removal Lemma of [Blum et al. 1998] and [Dunagan and Vempala 2001].

THEOREM 18. *Assume that at least an α probability mass of examples x have margin $\frac{|x \cdot c^*|}{|x||c^*|} \geq \gamma$ with respect to the target separator c^*. There is a polynomial-time algorithm (polynomial in d, $1/\gamma$, $1/\alpha$, $1/\epsilon$, and $1/\delta$) to learn a linear separator under the above assumptions, from a polynomial number of unlabeled examples and a single labeled example.*

PROOF. We prove the result by first arguing that a *random* halfspace has at least a $poly(\alpha, \gamma)$ probability of being a weak predictor. ([Blum et al. 1998] uses the Perceptron algorithm to get weak learning; here, we need something simpler since we need to save our labeled example to the very end.) Specifically, consider a point x of margin $\gamma_x \geq \gamma$. By definition, the margin is the cosine of the angle between x and c^*, and therefore the angle between x and c^* is $\pi/2 - \cos^{-1}(\gamma_x) \leq \pi/2 - \gamma$. Now, imagine that we draw f at random subject to $f \cdot c^* \geq 0$ (half of the f's will have this property) and define $f(x) = \text{sign}(f \cdot x)$. Then,

$$\mathbf{Pr}_f(f(x) \neq c^*(x)|f \cdot c^* \geq 0) \leq (\pi/2 - \gamma)/\pi = 1/2 - \gamma/\pi.$$

Moreover, if x does *not* have margin γ then at the very least we have $\mathbf{Pr}_f(f(x) \neq c^*(x)|f \cdot c^* \geq 0) \leq 1/2$. So, overall, since at least an α fraction of the points have

margin at least γ, we have

$$\mathbf{E}_f[err(f)|f \cdot c^* \geq 0] \leq 1/2 - \alpha\gamma/\pi.$$

Since $err(f)$ is a bounded quantity, this means that a $poly(\alpha, \gamma)$ probability mass of functions f must in fact be weakly-useful predictors.

The second step of the algorithm is as follows. Using the above observation, we pick a random f, and plug it into the bootstrapping theorem of [Blum and Mitchell 1998] (which, given a distribution over unlabeled pairs $\langle x_1^i, x_2^i \rangle$, will use $f(x_1^i)$ as a noisy label of x_2^i, feeding the result into a Statistical Query algorithm), repeating this process $poly(1/\alpha, 1/\gamma, \log(1/\delta))$ times. With high probability, our random f was a weakly-useful predictor on at least one of these steps, and we end up with a low-error hypothesis. For the rest of the runs of the algorithm, we have no guarantees. We now observe the following. First of all, any function f with small $err(f)$ must have small $err_{unl}(f)$; in particular, $\mathbf{Pr}(f(x_1) \neq f(x_2)) \leq 2err(f)$. Secondly, because of the assumption of independence given the label, given an unlabeled sample of size $poly(1/\tau, d, 1/\delta)$, as shown in Theorem 15, with high probability the *only* functions with unlabeled error at most τ are functions 2τ-close to c^*, 2τ-close to $\neg c^*$, 2τ-close to the "all positive" function, or 2τ-close to the "all negative" function.

We now simply examine the hypotheses produced by this procedure, and pick some f with a low empirical unlabeled error rate that is empirically at least 3τ-far from the "all-positive" or "all-negative" functions (if no such f exists, then this means the target must be 6τ-close to the all-positive or all-negative function so we simply choose $f =$ "all positive"). By the above argument, with high probability either f or $\neg f$ is 6τ-close to c^*. We can now just use $\mathcal{O}\left(\log_{\left(\frac{1}{\tau}\right)}\left(\frac{1}{\delta}\right)\right)$ labeled examples to determine which case is which (Lemma 16). This quantity is at most 1 and our error rate is at most ϵ if we set $\tau \leq \epsilon/6$ and τ sufficiently small compared to δ. This completes the proof. □

The above algorithm assumes one can efficiently pick a random unit-length vector in R^d, but the argument easily goes through even if we do this to only $O(\log 1/\gamma)$ bits of precision.

We now extend the result to the case that we make no margin assumption.

THEOREM 19. *There is a polynomial-time algorithm (in d, b, $1/\epsilon$, and $1/\delta$, where d is the dimension of the space and b is the number of bits per example) to learn a linear separator under the above assumptions, from a polynomial number of unlabeled examples and a single labeled example. Thus, we efficiently* PAC_{SSL}*-learn the class of linear separators over $\{-2^b, \ldots, 2^b - 1, 2^b\}^d$ under the agreement notion of compatibility if the distribution D satisfies independence given the label.*

PROOF. We begin by drawing a large unlabeled sample S (of size polynomial in d and b). We then compute a linear transformation T that when applied to S has the property that for any hyperplane $w \cdot x = 0$, at least a $1/poly(d, b)$ fraction of $T(S)$ has margin at least $1/poly(d, b)$. We can do this via the Outlier Removal Lemma of [Blum et al. 1998] and [Dunagan and Vempala 2001]. Specifically, the Outlier Removal Lemma states that given a set of points S, one can algorithmically remove an ϵ' fraction of S and ensure that for the remaining set S', for any vector w,

$\max_{x \in S'}(w \cdot x)^2 \leq poly(d, b, 1/\epsilon')\mathbf{E}_{x \in S'}[(w \cdot x)^2]$, where b is the number of bits needed to describe the input points. Given such a set S', one can then use its eigenvectors to compute a standard linear transformation (also described in [Blum et al. 1998]) $T : R^d \to R^{d'}$, where $d' \leq d$ is the dimension of the subspace spanned by S', such that in the transformed space, for all unit-length w, we have $\mathbf{E}_{x \in T(S')}[(w \cdot x)^2] = 1$. In particular, since the maximum of $(w \cdot x)^2$ is bounded, this implies that for any vector $w \in R^{d'}$, at least an α fraction of points $x \in T(S')$ have margin at least α for some $\alpha \geq 1/poly(b, d, 1/\epsilon')$.

Now, choose $\epsilon' = \epsilon/4$, and let D' be the distribution $\tilde{D}$ restricted to the space spanned by S'. By VC-dimension bounds, $|S| = \tilde{O}(d/\alpha)$ is sufficient so that with high probability, (a) D' has probability mass at least $1 - \epsilon/2$, and (b) the vector $T(c^*)$ has at least an $\alpha/2$ probability mass of $T(D')$ at margin $\geq \alpha$. Thus, the linear transformation T converts the distribution D' into one satisfying the conditions needed for Theorem 18, and any hypothesis produced with error $\leq \epsilon/2$ on D' will have error at most ϵ on D. So, we simply apply T to D' and run the algorithm for Theorem 18 to produce a low-error linear separator. □

Note: We can easily extend our algorithm to the standard co-training setting (where c_1^* can be different from c_2^*) as follows: we repeat the procedure in a symmetric fashion, and then just try all combinations of pairs of functions returned to find one of small unlabeled error rate, not close to "all positive", or "all negative". Finally we use $\mathcal{O}\left(\log_{\left(\frac{1}{\epsilon}\right)}\left(\frac{1}{\delta}\right)\right)$ labeled examples to produce a low error hypothesis (and here we use only one part of the example and only one of the functions in the pair).

5. RELATED MODELS

In this section we discuss a transductive analog of our model, some connections with generative models and other ways of using unlabeled data in Machine Learning, and the relationship between our model and the luckiness framework of [Shawe-Taylor et al. 1998].

5.1 A Transductive Analog of our Model

In *transductive* learning, one is given a fixed set S of examples, of which some small random subset is labeled, and the goal is to predict well on the rest of S. That is, we know which examples we will be tested on up front, and in a sense this a case of learning from a known distribution (the uniform distribution over S). We can also talk about a transductive analog of our inductive model, that incorporates many of the transductive learning methods that have been developed. In order to make use of unlabeled examples, we will again express the relationship we hope the target function has with the data through a compatibility notion χ. However, since in this case the compatibility of a given hypothesis is completely determined by S (which is known), we will not need to require that compatibility be an expectation over unlabeled examples. From the sample complexity point of view we only care about how much labeled data we need, and algorithmically we need to find a highly compatible hypothesis with low error on the labeled data.

Rather than presenting general theorems, we instead focus on the modeling question, and show how a number of existing transductive graph-based learning algo-

rithms can be modeled in our framework. In these methods one usually assumes that there is weighted graph G defined over S, which is given a-priori and encodes the prior knowledge. In the following we denote by W the weighted adjacency matrix of G and by $\mathcal{C}_S$ the set of all binary functions over S.

Minimum cut. Suppose for $f \in \mathcal{C}_S$ we define the incompatibility of f to be the weight of the cut in G determined by f. This is the implicit notion of compatibility considered in [Blum and Chawla 2001], and algorithmically the goal is to find the most compatible hypothesis that is correct on the labeled data, which can be solved efficiently using network flow. From a sample-complexity point of view, the number of labeled examples we need is proportional to the VC-dimension of the class of hypotheses that are at least as compatible as the target function. This is known to be $\mathcal{O}\left(\frac{k}{\lambda}\right)$ [Kleinberg 2000; Kleinberg et al. 2004], where k is the number of edges cut by c^* and λ is the size of the global minimum cut in the graph. Also note that the Randomized Mincut algorithm (considered by [Blum et al. 2004]), which is an extension of the basic mincut approach, can be viewed as motivated by a PAC-Bayes sample complexity analysis of the problem.

Normalized Cut. For $f \in \mathcal{C}_S$ define $size(f)$ to be the weight of the cut in G determined by f, and let $neg(f)$ and $pos(f)$ be the number of points in S on which f predicts negative and positive, respectively. For the normalized cut setting of [Joachims 2003] we can define the incompatibility of $f \in \mathcal{C}_S$ to be $\frac{size(f)}{neg(f) \cdot pos(f)}$. This is the penalty function used in [Joachims 2003], and again, algorithmically the goal would be to find a highly compatible hypothesis that is correct on the labeled data. Unfortunately, the corresponding optimization problem is in this case is NP-hard. Still, several approximate solutions have been considered, leading to different semi-supervised learning algorithms. For instance, Joachims [2003] considers a spectral relaxation that leads to the "SGT algorithm"; another relaxation based on semidefinite programming is considered in [Bie and Cristianini 2004].

Harmonic Function. We can also model the algorithms introduced in [Zhu et al. 2003c; 2003a] as follows. If we consider f to be a probabilistic prediction function defined over S, then we can define the incompatibility of f to be

$$\sum_{i,j} w_{i,j} \left(f(i) - f(j)\right)^2 = f^T L f,$$

where L is the un-normalized Laplacian of G. Similarly we can model the algorithm introduced by Zhao et al. [Zhou et al. 2004] by using an incompatibility of f given by $f^T \mathcal{L} f$ where $\mathcal{L}$ is the normalized Laplacian of G. More generally, all the Graph Kernel methods can be viewed in our framework if we consider that the incompatibility of f is given by $||f||_K = f^T K f$ where K is a kernel derived from the graph (see for instance [Zhu et al. 2003b]).

5.2 Connections to Generative Models

It is also interesting to consider how generative models can be fit into our model. As mentioned in Section 1, a typical assumption in a generative setting is that D is a mixture with the probability density function $p(x|\theta) = p_0 \cdot p_0(x|\theta_0) + p_1 \cdot p_1(x|\theta_1)$ (see for instance [Ratsaby and Venkatesh 1995; Castelli and Cover 1995; 1996]). In other words, the labeled examples are generated according to the following

mechanism: a label $y \in \{0,1\}$ is drawn according to the distribution of classes $\{p_0, p_1\}$ and then a corresponding random feature vector is drawn according to the class-conditional density p_y. The assumption typically used is that the mixture is identifiable. Identifiability ensures that the Bayes optimal decision border $\{x : p_0 \cdot p_0(x|\theta_0) = p_1 \cdot p_1(x|\theta_1)\}$ can be deduced if $p(x|\theta)$ is known, and therefore one can construct an estimate of the Bayes border by using $p(x|\hat{\theta})$ instead of $p(x|\theta)$. Essentially once the decision border is estimated, a small labeled sample suffices to learn (with high confidence and small error) the appropriate class labels associated with the two disjoint regions generated by the estimate of the Bayes decision border. To see how we can incorporate this setting in our model, consider for illustration the setting in [Ratsaby and Venkatesh 1995]; there they assume that $p_0 = p_1$, and that the class conditional densities are d-dimensional Gaussians with unit covariance and unknown mean vectors $\theta_i \in R^d$. The algorithm used is the following: the unknown parameter vector $\theta = (\theta_0, \theta_1)$ is estimated from unlabeled data using a maximum likelihood estimate; this determines a hypothesis which is a linear separator that passes through the point $(\hat{\theta}_0 + \hat{\theta}_1)/2$ and is orthogonal to the vector $\hat{\theta}_1 - \hat{\theta}_0$; finally each of the two decision regions separated by the hyperplane is labeled according to the majority of the labeled examples in the region. Given this setting, a natural notion of compatibility we can consider is the expected log-likelihood function (where the expectation is taken with respect to the unknown distribution specified by θ). Specifically, we can identify a legal hypothesis $f_{\overline{\theta}}$ with the set of parameters $\overline{\theta} = (\overline{\theta}_0, \overline{\theta}_1)$ that determine it, and then we can define $\chi(f_{\overline{\theta}}, D) = \mathbf{E}_{x \in D}[\log(p(x|\overline{\theta}))]$. [Ratsaby and Venkatesh 1995] show that if the unlabeled sample is large enough, then all hypotheses specified by parameters $\overline{\theta}$ which are close enough to θ, will have the property that their empirical compatibilities will be close enough to their true compatibilities. This then implies (together with other observations about Gaussian mixtures) that the maximum likelihood estimate will be close enough to θ, up to permutations. (This actually motivates χ as a good compatibility function in our model.)

More generally, we can deal with other parametric families using the same compatibility notion; however, we will need to impose constraints on the distributions allowed in order to ensure that the compatibility is actually well defined (the expected log-likelihood is bounded).

As mentioned in Section 1, this kind of generative setting is really at the extreme of our model. The assumption that the distribution that generates the data is truly a mixture implies that if we knew the distribution, then there are only two possible concepts left (and this makes the unlabeled data extremely useful).

5.3 Connections to the Luckiness Framework

It is worth noticing that there is a strong connection between our approach and the luckiness framework [Shawe-Taylor et al. 1998; Mendelson and Philips 2003]. In both cases, the idea is to define an ordering of hypotheses that depends on the data, in the hope that we will be "lucky" and find that the target function appears early in the ordering. There are two main differences, however. The first is that the luckiness framework (because it was designed for supervised learning only) uses labeled data both for estimating compatibility and for learning: this is a

more difficult task, and as a result our bounds on labeled data can be significantly better. For instance, in Example 4 described in Section 2, for any non-degenerate distribution, a dataset of $\frac{d}{2}$ pairs can with probability 1 be completely shattered by fully-compatible hypotheses, so the luckiness framework does not help. In contrast, with a larger (unlabeled) sample, one can potentially reduce the space of compatible functions quite significantly, and learn from $o(d)$ or even $\mathcal{O}(1)$ labeled examples depending on the distribution – see Section 3.2 and Section 4. Secondly, the luckiness framework talks about compatibility between a hypothesis and a *sample*, whereas we define compatibility with respect to a distribution. This allows us to talk about the amount of unlabeled data needed to estimate true compatibility. There are also a number of differences at the technical level of the definitions.

5.4 Relationship to Other Ways of Using Unlabeled Data for Learning

It is well known that when learning under an unknown distribution, unlabeled data might help somewhat even in the standard discriminative models by allowing one to use both distribution-specific algorithms [Benedek and Itai 1991], [Kaariainen 2005], [Sokolovska et al. 2008] and/or tighter data dependent sample-complexity bounds [Bartlett and Mendelson 2002; Koltchinskii 2001]. However in all these methods one chooses a class of functions or a prior over functions *before* performing the inference. This does not capture the power of unlabeled data in many of the practical semi-supervised learning methods, where typically one has some idea about what structure of the data tells about the target function, and where the choice of prior can be made more precise after seeing the unlabeled data [Blum and Mitchell 1998; Joachims 1999; 1999; Leskes 2005; Rosenberg and Bartlett 2007]. Our focus in this work has been to provide a unified discriminative framework for reasoning about usefulness of unlabeled data in such settings in which one can analyze both sample complexity and algorithmic results.

6. CONCLUSIONS

The formulation of the PAC learning model by Valiant [1984] and the Statistical Learning Theory framework by Vapnik [1982] were instrumental in the development of machine learning and the design and analysis of algorithms for supervised learning. Many modern learning problems, however, call for *semi-supervised* methods that can take advantage of large quantities of unlabeled data that are often available, and while a large number of algorithms have been explored, there has been no unifying theoretical framework. In this paper, we develop such a framework that captures many of the ways unlabeled data is typically used, and the fundamental assumptions underlying these approaches. This framework allows one to analyze when and why unlabeled data can help and what the basic quantities are that these data bounds depend on. The high level implication of our analysis is that unlabeled data is useful if (a) we have a good notion of compatibility so that the target function indeed has a low unlabeled error rate, (b) the distribution D is *helpful* in the sense that not too many other hypotheses also have a low unlabeled error rate, and (c) we have enough *unlabeled* data to estimate unlabeled error rates well. We then make these statements precise through a series of sample-complexity results, giving bounds as well as identifying the key quantities of interest. In addition, we

give several efficient algorithms for learning in this framework. One consequence of our model is that if the target function and data distribution are both well behaved with respect to the compatibility notion, then the sample-size bounds we get can substantially beat what one could hope to achieve using labeled data alone, and we have illustrated this with a number of examples through the paper.

6.1 Subsequent Work

Following the initial publication of this work, several authors have used our framework for reasoning about semi-supervised learning, as well as for developing new algorithms and analyses of semi-supervised learning. For example [Shawe-Taylor 2006; Rosenberg and Bartlett 2007; Ganchev et al. 2008] use it in the context of agreement-based multi-view learning for either classification with specific convex loss functions (e.g., hinge loss) or for regression. Sridharan and Kakade [2008] use our framework in order to provide a general analysis multi-view learning for a variety of loss functions and learning tasks (classification and regression) along with characterizations of suitable notions of compatibility functions. Parts of this work appear as a book chapter in [Chapelle et al. 2006] and as stated in the introduction of that book, our framework can be used to obtain bounds for a number of the semi-supervised learning methods used in the other chapters.

6.2 Open Problems and Future Directions

Our work brings up a number of open questions, both specific and high-level. One broad category of such questions is for what natural classes $\mathcal{C}$ and compatibility notions χ can one provide an efficient algorithm that PAC_{SSL}-learns the pair $(\mathcal{C}, \chi)$: i.e., an algorithm whose running time and sample sizes are polynomial in the bounds of Theorem 4? For example, a natural question of this form is: can one generalize the algorithm of Section 4.1 to allow for irrelevant variables that are neither positive nor negative indicators? That is, suppose we define a "two-sided disjunction" h to be a pair of disjunctions (h_+, h_-) where h is compatible with D iff for all examples x, $h_+(x) = -h_-(x)$ (and let us define $h(x) = h_+(x)$). Can we efficiently learn the class of two-sided disjunctions under this notion of compatibility?

Alternatively, as a different generalization of the problem analyzed in Section 4.1, suppose that again every variable is either a positive or negative indicator, but we relax the "margin" condition. In particular, suppose we require that every example x either contain at least 60% of the positive indicators and at most 40% of the negative indicators (for positive examples) or vice versa (for negative examples). Can this class be learned efficiently with bounds comparable to those from Theorem 4? Along somewhat different lines, can one generalize the algorithm given for Co-Training with linear separators, to assume some condition weaker than independence given the label, while maintaining computational efficiency?

More broadly, it would be interesting to extend our model to related settings such as that of *active learning*. As in semi-supervised learning, in active learning the algorithm initially sees only the unlabeled portion of a pool of examples drawn from some underlying distribution. However, in active learning, the algorithm then gets to decide which examples in the pool to have labeled for it rather than just getting labels for a random subset of examples. Thus, potentially many fewer labelings may be necessary. There have recently been a number of papers and algorithmic

results (both theoretical and practical) for very special settings (see e.g. [Balcan et al. 2007; Castro and Nowak 2007; Hanneke 2007; Balcan et al. 2008; Dasgupta et al. 2007; Balcan et al. 2006; Kääriäinen 2006; Dasgupta 2005; 2004; Freund et al. 1993]), but a more complete understanding is still missing.

Our framework can be viewed as falling under the general area of learning with data-dependent hypothesis spaces, and it would also be interesting to analyze this for related settings such as learning with weakly-labeled data or additional information from world knowledge as in [Gabrilovich and Markovitch 2005].

REFERENCES

Amini, M.-R., Chapelle, O., and Ghani, R., Eds. 2005. *Learning with Partially Classified Training Data*. Workshop, ICML'05.

Balcan, M. F., Beygelzimer, A., and Langford, J. 2006. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*.

Balcan, M.-F., Blum, A., and Yang, K. 2004. Co-training and expansion: Towards bridging theory and practice. In *Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems*.

Balcan, M.-F., Broder, A., and Zhang, T. 2007. Margin based active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*.

Balcan, M.-F., Hanneke, S., and Wortman, J. 2008. The true sample complexity of active learning. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*.

Bartlett, P., Boucheron, S., and Lugosi, G. 1999. Model selection and error estimation. In *Proceedings of the 13th Annual Conference on Computational Learning Theory*.

Bartlett, P. and Mendelson, S. 2002. Rademacher and gaussian complexities risk bounds and structural results. *Journal of Machine Learning Research*, 463–482.

Baum, E. B. 1990. Polynomial time algorithms for learning neural nets. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*. 258 – 272.

Benedek, G. and Itai, A. 1991. Learnability with respect to a fixed distribution. *Theoretical Computer Science 86*, 377–389.

Bie, T. D. and Cristianini, N. 2003. Convex methods for transduction. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems*. Vol. 16.

Bie, T. D. and Cristianini, N. 2004. Convex transduction with the normalized cut. Internal Report 04-128, ESAT-SISTA, K.U.Leuven.

Blum, A. and Chawla, S. 2001. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Blum, A., Frieze, A., Kannan, R., and Vempala, S. 1998. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica 22*, 35–52.

Blum, A. and Kannan, R. 1997. Learning an intersection of k halfspaces over a uniform distribution. *Journal of Computer and Systems Sciences 54,* 2, 371–380.

Blum, A., Lafferty, J., Reddy, R., and Rwebangira, M. R. 2004. Semi-supervised learning using randomized mincuts. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

Blum, A. and Mitchell, T. M. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. 92–100.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. 1989. Learnability and the Vapnik Chervonenkis dimension. *Journal of the ACM 36(4)*, 929–965.

Boucheron, S., Bousquet, O., and Lugosi, G. 2005. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics 9*, 9:323–375.

Boucheron, S., Lugosi, G., and Massart, P. 2000. A sharp concentration inequality with applications. *Random Structures and Algorithms 16*, 277–292.

Castelli, V. and Cover, T. 1995. On the exponential value of labeled samples. *Pattern Recognition Letters 16*, 105–111.

Castelli, V. and Cover, T. 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory 42(6)*, 2102–2117.

Castro, R. and Nowak, R. 2007. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*.

Chapelle, O., Schölkopf, B., and Zien, A., Eds. 2006. *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Chapelle, O. and Zien, A. 2005. Semi-supervised classification by low density separation. In *Tenth International Workshop on Artificial Intelligence and Statistics*.

Collins, M. and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 189–196.

Dasgupta, S. 2004. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems (NIPS)*.

Dasgupta, S. 2005. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*. Vol. 18.

Dasgupta, S., Hsu, D., and Monteleoni, C. 2007. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems 20*.

Dasgupta, S., Littman, M. L., and McAllester, D. 2001. Pac generalization bounds for co-training. In *Advances in Neural Information Processing Systems 14*.

Devroye, L., Gyorfi, L., and Lugosi, G. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

Dunagan, J. and Vempala, S. 2001. Optimal outlier removal in high-dimensional spaces. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. 1989. A general lower bound on the number of examples needed for learning. *Information and Computation 82*, 246–261.

Flaxman, A. 2003. Personal communication.

Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. 1993. Information, prediction, and query by comittee. In *Neural Information Processing Systems*.

Gabrilovich, E. and Markovitch, S. 2005. Feature generation for text categorization using world knowledge. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*. 1048–1053.

Ganchev, K., Graca, J., Blitzer, J., and Taskar, B. 2008. Multi-view learning over structured and non-identical outputs. In *Proceedings of The 24th Conference on Uncertainty in Artificial Intelligence*.

Ghani, R. 2001. Combining labeled and unlabeled data for text classification with a large number of categories. In *Proceedings of the IEEE International Conference on Data Mining*.

Ghani, R., Jones, R., and Rosenberg, C., Eds. 2003. *The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*. Workshop, ICML'03.

Hanneke, S. 2007. A bound on the label complexity of agnostic activelearning. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML)*.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*. 200–209.

Joachims, T. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning*.

Kaariainen, M. 2005. Generalization error bounds using unlabeled data. In *Proceedings of the 18th Annual Conference on Learning Theory*. 127–142.

Kääriäinen, M. 2006. On active learning in the non-realizable case. In *Proceedings of 17th International Conference on Algorithmic Learning Theory (ALT)*. Lecture Notes in Computer Science, vol. 4264. 63–77.

Kearns, M. 1998. Efficient noise-tolerant learning from statistical queries. In *Journal of the ACM (JACM)*. 983 – 1006.

Kearns, M. and Vazirani, U. 1994. *An Introduction to Computational Learning Theory*. MIT Press.

Kleinberg, J. 2000. Detecting a network failure. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*. 231–239.

Kleinberg, J., Sandler, M., and Slivkins, A. 2004. Network failure detection and graph connectivity. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*. 231–239.

Klivans, A. R., O'Donnell, R., and Servedio, R. 2002. Learning intersections and thresholds of halfspaces. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*. 177–186.

Koltchinskii, V. 2001. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory 47,* 5, 1902–1914.

Leskes, B. 2005. The value of agreement, a new boosting algorithm. In *Proceedings of the Eighteenth Annual Conference on Learning Theory*. 51 – 56.

Levin, A., Viola, P., and Freund, Y. 2003. Unsupervised improvement of visual detectors using co-training. In *Proceedings of the Nineth IEEE International Conference on Computer Vision*. 626–633.

Linial, N., Mansour, Y., and Nisan, N. 1989. Constant depth circuits, fourier transform, and learnability. In *Proceedings of the Thirtieth Annual Symposium on Foundations of Computer Science*. 574–579.

Mendelson, S. and Philips, P. 2003. Random subclass bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*.

Nigam, K. and Ghani, R. 2000. Analyzing the effectiveness and applicability of co-training. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Management*. 86–93.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. M. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning 39(2/3)*, 103–134.

Park, S. and Zhang, B. 2003. Large scale unstructured document classification using unlabeled data and syntactic information. In *PAKDD 2003*. LNCS vol. 2637. Springer, 88–99.

Pierce, D. and Cardie, C. 2001. Limitations of Co-Training for natural language learning from large datasets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1–9.

Ratsaby, J. and Venkatesh, S. 1995. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*. 412–417.

Rosenberg, D. and Bartlett, P. 2007. The Rademacher Complexity of Co-Regularized Kernel Classes. In *Proceedings of Artificial Intelligence & Statistics*.

Shawe-Taylor, J. 2006. Rademacher Analysis and Multi-View Classification. http://www.gla.ac.uk/external/RSS/RSScomp/shawe-taylor.pdf.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. 1998. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory 44(5)*, 1926–1940.

Sokolovska, N., Capp, O., and Yvon, F. 2008. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th International Conference on Machine Learning*.

Sridharan, K. and Kakade, S. M. 2008. An information theoretic framework for multi-view learning. In *Proceedings of the 21st Annual Conference on Learning Theory*.

Valiant, L. 1984. A theory of the learnable. *Communications of the ACM 27(11)*, 1134–1142.

Vapnik, V. N. 1982. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York.

Vapnik, V. N. 1998. *Statistical Learning Theory*. John Wiley and Sons Inc.

Vempala, S. 1997. A random sampling based algorithm for learning the intersection of half-spaces. In *Proceedings of the 38th Symposium on Foundations of Computer Science*. 508–513.

Verbeurgt, K. A. 1990. Learning dnf under the uniform distribution in quasi-polynomial time. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*. 314–326.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*. 189–196.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schlkopf, B. 2004. Learning with local and global consistency. In *Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems*.

Zhu, X. 2006. Semi-Supervised Learning Literature Survey. Computer Sciences TR 1530 University of Wisconsin - Madison.

Zhu, X., Ghahramani, Z., and Lafferty, J. 2003a. Combinig active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning*. Washington, DC, USA, 912–912.

Zhu, X., Ghahramani, Z., and Lafferty, J. 2003b. Semi-supervised learning: From gaussian fields to gaussian processes. Tech. rep., Carnegie Mellon University.

Zhu, X., Ghahramani, Z., and Lafferty, J. 2003c. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of The Twentieth International Conference on Machine Learning*. 912–912.

A. STANDARD RESULTS

In this appendix we state a few known generalization bounds and concentration results used in our proofs. We start with several classic results. See, e.g., [Devroye et al. 1996].

Theorem 20. *Suppose that $\mathcal{C}$ is a set of functions from X to $\{0,1\}$ with finite VC-dimension $V \geq 1$. For any distribution D over X, any target function (not necessarily in $\mathcal{C}$), and any ϵ, $\delta > 0$, if we draw a sample from D of size*

$$m(\epsilon, \delta, V) = \frac{64}{\epsilon^2}\left(2V \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right),$$

then with probability at least $1-\delta$, we have $|err(h) - \widehat{err}(h)| \leq \epsilon$ for all $f \in \mathcal{C}$.

Theorem 21. *Suppose that $\mathcal{C}$ is a set of functions from X to $\{0,1\}$ with finite VC-dimension $V \geq 1$. For any probability distribution D over X, any target function c^*, we have*

$$\mathbf{Pr}_S\left[\sup_{f \in C, \widehat{err}(f)=0} |err(f) - \widehat{err}(f)| \geq \epsilon\right] \leq 2\mathcal{C}[2m, D]e^{-m\epsilon/2}.$$

So, for any ϵ, $\delta > 0$, if we draw a sample from D of size

$$m \geq \frac{2}{\epsilon}\left(2\ln\left(\mathcal{C}[2m, D]\right) + \ln\left(\frac{2}{\delta}\right)\right),$$

then with probability at least $1-\delta$, we have that all functions with $\widehat{err}(f) = 0$ satisfy $err(f) \leq \epsilon$.

Theorem 22. *Suppose that $\mathcal{C}$ is a set of functions from X to $\{0,1\}$ with finite VC-dimension $V \geq 1$. For any probability distribution D over X, any target*

function c^*, we have

$$\mathbf{Pr}_S\left[\sup_{f\in C}|err(f)-\widehat{err}(f)|\geq\epsilon\right]\ \leq\ 8\mathcal{C}[2m,D]e^{-m\epsilon^2/8}.$$

So, for any ϵ, $\delta>0$, if we draw from D a sample satisfying

$$m\geq\frac{8}{\epsilon^2}\left(\ln\left(\mathcal{C}[m,D]\right)+\ln\left(\frac{8}{\delta}\right)\right),$$

then with probability at least $1-\delta$ all functions f satisfy $|err(f)-\widehat{err}(f)|\leq\epsilon$.

We now state a result from [Boucheron et al. 2000].

THEOREM 23. *Suppose that $\mathcal{C}$ is a set of functions from X to $\{0,1\}$. Let D be an arbitrary, but fixed probability distribution over X. For any target function and for any i.i.d. sample of S of size m from D, let f_m be the function that minimizes the empirical error over S. Then for any $\delta>0$, the probability that*

$$err(f_m)\leq\widehat{err}(f_m)+\sqrt{\frac{6\ln\mathcal{C}[S]}{m}}+4\sqrt{\frac{\ln(2/\delta)}{m}}$$

is greater than $1-\delta$.

Note that in fact the above statement is true even if in the right-hand side we use $C[S']$ instead of $C[S]$ where S' is another i.i.d sample of size m drawn from D.

THEOREM 24. *For any class of functions we have:*

$$\mathbf{Pr}_S[\log_2(\mathcal{C}[S])\geq\mathbf{E}[\log_2(\mathcal{C}[S])]+\alpha]\ \leq\ \exp\left[-\frac{\alpha^2}{2\mathbf{E}[\log_2(\mathcal{C}[S])]+2\alpha/3}\right].\quad(2)$$

Also,

$$\mathbf{E}[\log_2\mathcal{C}[S]]\leq\log_2\mathbf{E}[\mathcal{C}[S]]\leq\frac{1}{\ln 2}\mathbf{E}[\log_2\mathcal{C}[S]].\quad(3)$$

B. ADDITIONAL PROOFS

LEMMA 16. Consider $\tau<\frac{1}{8}$. Let $C_\tau=\left\{f,\overline{f}\right\}$ be a subset of $\mathcal{C}$ containing two opposite hypotheses with the property that one of them is τ-close to c^*. Then, $m_l>6\log_{\left(\frac{1}{\tau}\right)}\left(\frac{1}{\delta}\right)$ labeled examples are sufficient so that with probability $\geq1-\delta$, the concept in C_τ that is τ-close to c^* in fact has lower empirical error.

PROOF. We need to show that if $m_l>6\log_{\frac{1}{\tau}}\left(\frac{1}{\delta}\right)$, then

$$\sum_{k=0}^{\lfloor\frac{m_l}{2}\rfloor}\binom{m_l}{k}\tau^{(m_l-k)}\left(1-\tau\right)^k\leq\delta.$$

Since $\tau<\frac{1}{8}$ we have:

$$\sum_{k=0}^{\lfloor\frac{m_l}{2}\rfloor}\binom{m_l}{k}\tau^{(m_l-k)}\left(1-\tau\right)^k\leq\sum_{k=0}^{\lfloor\frac{m_l}{2}\rfloor}\binom{m_l}{k}\tau^{(m_l-k)}=\tau^{\lfloor\frac{m_l}{2}\rfloor}\sum_{k=0}^{\lfloor\frac{m_l}{2}\rfloor}\binom{m_l}{k}\tau^{\lceil\frac{m_l}{2}\rceil-k}$$

and so $S\leq\left(\sqrt{\tau}\cdot2\right)^{m_l}$. For $\tau<\frac{1}{8}$ and $m_l>6\frac{\log_2\left(\frac{1}{\delta}\right)}{\log_2\left(\frac{1}{\tau}\right)}=6\log_{\left(\frac{1}{\tau}\right)}\left(\frac{1}{\delta}\right)$ it's easy to see that $\left(\sqrt{\tau}\cdot2\right)^{m_l}<\delta$, which implies the desired result. □

THEOREM 25. *For any class of functions we have:*

$$\mathbf{Pr}_S[\log_2(\mathcal{C}[S]) \geq 2\log \mathbf{E}[\mathcal{C}[S]] + \alpha] \ \leq \ e^{-2\alpha}. \tag{4}$$

PROOF. Inequality (2) implies that:

$$\mathbf{Pr}_S[\log_2(\mathcal{C}[S]) \geq 2\mathbf{E}[\log_2(\mathcal{C}[S])] + \alpha] \leq \exp\left[-\frac{(\alpha + \mathbf{E}[\log_2(\mathcal{C}[S])])^2}{2\mathbf{E}[\log_2(\mathcal{C}[S])] + 2(\mathbf{E}[\log_2(\mathcal{C}[S])] + \alpha)/3}\right].$$

Since $\frac{(\alpha+a)^2}{2a+2(a+\alpha)/3} \geq \frac{\alpha}{2}$ for any $a \geq 0$ we get

$$\mathbf{Pr}_S[\log_2(\mathcal{C}[S]) \geq 2\mathbf{E}[\log_2(\mathcal{C}[S])] + \alpha] \leq e^{-\alpha/2}.$$

Combining this together with the following fact (implied by Inequality (3))

$$\mathbf{Pr}_S[\log_2(\mathcal{C}[S]) \geq 2\log \mathbf{E}[\mathcal{C}[S]] + \alpha] \leq \mathbf{Pr}_S[\log_2(\mathcal{C}[S]) \geq 2\mathbf{E}[\log_2(\mathcal{C}[S])] + \alpha],$$

we get the desired result. □