Fundamental Limitations of Semi-Supervised Learning

by

Tyler (Tian) Lu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2009

Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The emergence of a new paradigm in machine learning known as *semi-supervised learning* (SSL) has seen benefits to many applications where labeled data is expensive to obtain. However, unlike supervised learning (SL), which enjoys a rich and deep theoretical foundation, semi-supervised learning, which uses additional unlabeled data for training, still remains a theoretical mystery lacking a sound fundamental understanding. The purpose of this research thesis is to take a first step towards bridging this theory-practice gap.

We focus on investigating the inherent limitations of the benefits semi-supervised learning can provide over supervised learning. We develop a framework under which one can analyze the potential benefits, as measured by the sample complexity of semi-supervised learning. Our framework is utopian in the sense that a semi-supervised algorithm trains on a labeled sample and an unlabeled distribution, as opposed to an unlabeled sample in the usual semi-supervised model. Thus, any lower bound on the sample complexity of semi-supervised learning in this model implies lower bounds in the usual model.

Roughly, our conclusion is that unless the learner is absolutely certain there is some non-trivial relationship between labels and the unlabeled distribution ("SSL type assumption"), semi-supervised learning cannot provide significant advantages over supervised learning. Technically speaking, we show that the sample complexity of SSL is no more than a constant factor better than SL for any unlabeled distribution, under a *no-prior-knowledge* setting (i.e. without SSL type assumptions).

We prove that for the class of thresholds in the realizable setting the sample complexity of SL is at most twice that of SSL. Also, we prove that in the agnostic setting for the classes of thresholds and union of intervals the sample complexity of SL is at most a constant factor larger than that of SSL. We conjecture this to be a general phenomenon applying to any hypothesis class.

We also discuss issues regarding SSL type assumptions, and in particular the popular cluster assumption. We give examples that show even in the most accommodating circumstances, learning under the cluster assumption can be hazardous and lead to prediction performance much worse than simply ignoring unlabeled data and doing supervised learning.

This thesis concludes with a look into future research directions that builds on our investigation.

Acknowledgements

I would like to thank my advisor, Shai Ben-David, for his support, encouragement, and guidance throughout my Master's studies. I have benefitted a great deal from my advisor's high standards of scholarship and intellectually engaging style.

It has been a pleasure to work with my collaborators and colleagues Alex López-Ortiz, Kate Larson, Dávid Pál, Martin Pál, Teresa Luu, Rita Ackerman, Miroslava Sotáková, Sharon Wulff, Pooyan Khajehpour, and Alejandro Salinger. In addition, I like thank my office mates Ke Deng, Jakub Gawryjolek, Ting Liu, Sharon Wulff, Derek Wong, and Teresa Luu for creating a livelier atmosphere at work.

I like to thank some faculty members including Pascal Poupart, Joseph Cheriyan, Peter van Beek, Anna Lubiw, Yuying Li, Jochen Konemann, and Ali Ghodsi for teaching great courses or having interesting discussions on various topics.

I like to thank Pascal Poupart and Ming Li for reading my thesis and providing valuable and insightful feedback.

Lastly I would like to thank my parents and my extended family for their unconditional support and hope in me.

Contents

List of Figures **vii**

1 Introduction **1**

1.1 Outline of Thesis 4

2 Background in Statistical Learning Theory **6**

2.1 Some Notation 6

2.2 Motivating the Probably Approximately Correct Model 8

2.3 The Agnostic PAC Model 10

2.4 The Realizable PAC Model 12

2.5 Learnability and Distribution-Free Uniform Convergence 13

2.5.1 Agnostic Setting 14

2.5.2 Realizable Setting 16

3 Modelling Semi-Supervised Learning **18**

3.1 Utopian Model of Semi-Supervised Learning 19

3.2 Related Work 21

3.2.1 Previous Theoretical Approaches 24

3.3 Issues with the Cluster Assumption 25

4 Inherent Limitations of Semi-Supervised Learning **30**

4.1 Fundamental Conjecture on No-Prior-Knowledge SSL 31

4.2 Reduction to the Uniform Distribution on $[0, 1]$ 34

4.2.1 The Rough Idea 35

4.2.2 Formal Proof 36

4.3 Learning Thresholds in the Realizable Setting 40

4.4 Thresholds and Union of Intervals in the Agnostic Setting 48

4.5 No Optimal Semi-Supervised Algorithm 54

5 Conclusion **56**

5.1 Proving Conjectures 4.1 and 4.2 57

5.2 Verifying SSL Assumptions 57

5.3 SSL Sample Complexity Incorporating Assumptions 58

APPENDICES **60**

A Examples of VC Dimension **61**

A.1 Thresholds 61

A.2 Union of Intervals 62

A.3 Linear Halfspaces 62

References **63**

List of Figures

1.1 Guarantees can be proved for supervised learning 2

1.2 Can guarantees be proved for semi-supervised learning? 3

2.1 Example of the overfitting phenomenon 9

3.1 Example of problems with the cluster assumption on mixture of two Gaussian with same shape. 27

3.2 Example of problems with the cluster assumption on mixture of two very different Gaussians 28

3.3 Example of problems with the cluster assumption on mixture of two homogeneous labeled distributions 28

4.1 Reducing to the Uniform Distribution on $[0, 1]$ 35

4.2 Illustrating the algorithm of Kääriäinen (2005) 41

4.3 A sample complexity lower bound for distinguishing two similar Bernoulli distributions. 50

4.4 Example of a b-shatterable triple $(\mathbb{R}^2, \mathcal{H}, D)$ 52

4.5 Union of at most d intervals $2d$-shatters the uniform distribution over $[0, 1]$ 54

A.1 Class of thresholds 61

“The beginning of knowledge is the discovery of something we do not understand.”

— Frank Herbert (1920 - 1986)

Chapter 1

Introduction

Machine learning is a field that is broadly concerned with designing computer algorithms that learn from experience or automatically discover useful patterns from datasets. Much theory and applied work have been focused in the area of supervised learning, where the goal is to approximate a ground truth function $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is some domain, and $\mathcal{Y}$ is some label class. For example, $\mathcal{X}$ can be the set of all digital pictures, $\mathcal{Y}$ a set of some persons, and f tells us who appears in the photograph. Of course, our purpose is for the machine to automatically "learn" f, that is, output a function approximating f, based on an input collection of examples, of say, pictures and their correct labels, known as the *training data*,

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, \qquad \text{each } x_i \in \mathcal{X},\ y_i \in \mathcal{Y}.$$

In supervised learning, the theoretical frameworks defining learnability, the corresponding mathematical results of what can be learned, and the tradeoffs between learning an accurate predictor and the available training data are reasonably well understood. This has been a major success in the theoretical analysis of machine learning—a field known as *computational learning theory*, and *statistical learning theory* when computational complexity issues are ignored. Many practically important learning tasks can be cast in a supervised learning framework. Examples include predicting a patient's risk of heart disease given her medical records, making decisions about giving loans, face recognition in photographs, and many other important applications. See Figure 1.1 for an example of a theoretically sound supervised learning paradigm.

Outside of supervised learning, however, our current theoretical understanding of two important areas known as *unsupervised learning* and *semi-supervised learning* (SSL) leaves a lot to be desired. Unsupervised learning is concerned with discovering meaningful structure in a raw dataset. This may include grouping similar data points together, known as *clustering*, or finding a low dimensional embedding of high dimensional input data that can help in future data prediction problems, known as *dimensionality reduction*.

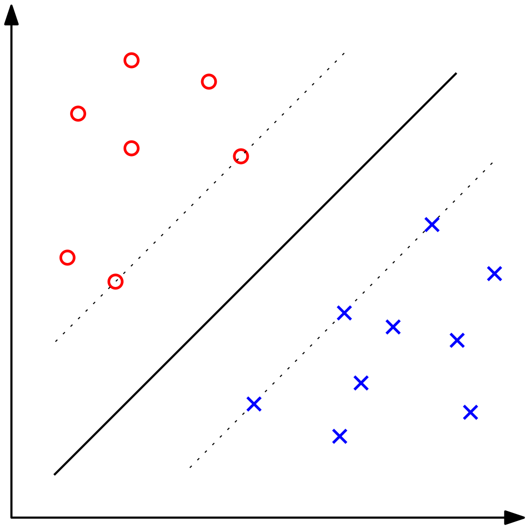

Figure 1.1: Supervised learning: a maximum margin linear classifier separating circles and crosses. Guarantees on future predictions can be given.

Semi-supervised learning, as the name suggests, is the task of producing a prediction rule given example data predictions (labeled data) and extra data without any prediction labels (unlabeled data). See Figure 1.2 for an illustration. In many practical scenarios, labeled data is expensive and hard to obtain as it usually requires human annotators to label the data, but unlabeled data are abundant and easily obtainable. Consequently, researchers are interested in using unlabeled data to help learn a better classifier. Due to numerous applications in areas such as natural language processing, bioinformatics, or computer vision, this use of auxiliary unlabeled data has been gaining attention in both the applied and theoretical machine learning communities. While semi-supervised learning heuristics abound (see for example, Zhu, 2008), the theoretical analysis of semi-supervised learning is distressingly scarce and does not provide a reasonable explanation of the advantages of unlabeled data.

While it may appear that learning with the addition of unlabeled data is magical—after all, what can one learn from data that does not provide any clue on a function that is to be approximated?—In fact most practitioners performing SSL make some type of assumptions on *how the labels behave with respect to the structure of the unlabeled data*. In practice, this may provide great advantage. For example, a popular assumption asserts that a good predictor should go through a low density region of the data domain. However, the focus of this thesis is on a more fundamental question: what can be gained when no assumptions of above type are made? While this question may seem far from practical interest, it is a first step towards the theoretical modelling of practical SSL and understanding its limitations.

In this thesis, we focus on formalizing a mathematical model of semi-supervised learning and analyze its potential benefits and inherent limitations when compared

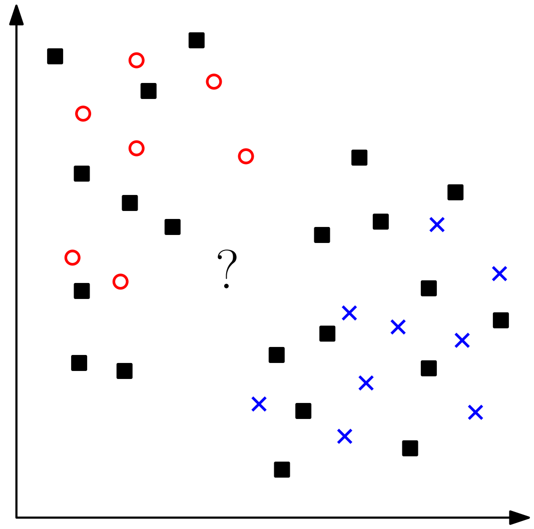

Figure 1.2: Semi-supervised learning: the extra boxes represent unlabeled data, but where to place the separator? What can be proved about the future performance?

with supervised learning. Our model is based on the Probably Approximately Correct (or PAC) learning framework proposed by Valiant (1984). The **main conclusion of our thesis is:**

> Unless the learner is *absolutely sure of an assumption that holds* on the relationship between labels and the unlabeled data structure (if no assumptions are made then we refer to it as the *no-prior-knowledge* setting) then one cannot hope to obtain a significant advantage in the sample complexity[1] of semi-supervised learning over that of supervised learning.

See Chapter 4 for a precise statement. When SSL assumptions are made but do not hold, it can degrade the performance and can be worse than supervised learning (see Section 3.3. The semi-supervised learning model used in our analysis is one in which the learner is provided with a labeled training sample and complete knowledge of the distribution generating unlabeled data. Of course, this differs from the real-world model where a sample of unlabeled data is given. However, our analysis shows that even in such an optimistic scenario that we assume, one still cannot obtain better than constant factor improvement in the *labeled* sample complexity. This is done by proving lower bounds on the labeled sample complexity of SSL, which also applies to supervised learning, and comparing that with the upper bounds on the labeled sample complexity of supervised learning, which also applies to SSL. In this thesis we are mainly concerned with lower bounds in our SSL model. At the

[1] Sample complexity refers to the amount of labeled training data needed to learn an accurate classifier. An alternate measure is the error rate of the learned classifier, which happens to depend on the sample complexity.

same time, upper bounds in our SSL model apply to the real-world SSL model as the unlabeled sample size grows.

We also show that common applications of SSL that assume some relationship between labels and the unlabeled data distribution (such as the widely held *cluster assumption* that prefers decision boundaries through low density region) may lead to poor prediction accuracy, even when the labeled data distribution does satisfy the assumption to a large degree (e.g. the data comes from a mixture of two Gaussian distributions, one for each class label).

Our thesis is not the first work on the merits of using unlabeled data without making assumptions regarding the relationship between labels and the unlabeled data distribution. The transductive learning model of Vapnik (2006) and its performance bounds do not make SSL type assumptions. However, the transductive model is concerned with prediction of a fixed, unlabeled test data set rather than generalizing a predictor for all points in a domain. As well, Kääriäinen (2005) proposes some SSL algorithms that do not depend on SSL type assumptions. Meanwhile, the work of Balcan and Blum (2005, 2006) offers a PAC style framework for formalizing SSL type assumptions. We discuss more about how our work compares with respect to these approaches and other related works in Section 3.2. The fundamental difference with our work is that we are interested in understanding the inherent limitations of the benefits that semi-supervised learning provides over supervised learning, whereas some of the related work is on providing "luckiness" conditions under which SSL can be successful.

This thesis does not completely provide answers to the question of the merits of semi-supervised learning. But it does show that for some relatively natural classes of prediction functions over the real line, semi-supervised learning does not help significantly unless additional assumptions are made. We also believe the results generalize to other classes of functions, as asserted in our conjectures in Section 4.1. Much of the results contained in this thesis has already appeared in preliminary form Ben-David et al. (2008).

1.1 Outline of Thesis

Chapter 2. We first present some background material on the statistical learning theory of supervised learning that will be essential to understanding our proposed formulation of semi-supervised learning in the later chapters. This chapter will start off with a brief expository tour of the main motivations and issues that must be addressed by a formal framework of learning, while presenting definitions along the way that will ultimately lead to constructing a formal framework of learning known as the Probably Approximately Correct (PAC) learning framework (Valiant, 1984). We also discuss the subtleties of PAC learning.

In the last part of the chapter, we review a seminal result of Vapnik and Chervonenkis (1971) on empirical process theory and its relationship in fully characterizing the informational requirements of PAC learning. Important theorems will be stated that provide upper and lower bounds on the sample complexity, in other words, how much training data is needed to learn well.

Chapter 3. In this chapter we will motivate and present a utopian model of semi-supervised learning as well as definitions corresponding to measuring its sample complexity. We will then briefly describe related work on SSL from the perspective of our work. We will also discuss in detail and critique previous theoretical paradigms for SSL including the shortcomings of these approaches. Then we turn our attention to an important issue for practitioners performing SSL, that of the potential hazards of learning under the popular *cluster assumption*. We give examples of scenarios that appear to be amenable for learning with the cluster assumption, but in actuality damages learning.

Chapter 4. We propose a fundamental conjecture under the no-prior-knowledge setting[2] that roughly asserts SSL cannot provide significant advantages over supervised learning. Then, for the remaining part of the chapter we turn our attention to proving the conjecture for some basic hypothesis classes over the real line.

For "natural" hypothesis classes over the real line, we present a reduction lemma that reduces semi-supervised learning under "smooth" distributions to supervised learning under the fixed uniform distribution on the unit interval, while preserving its sample complexity. Using this lemma, we are able to prove the conjecture for the class of thresholds in the realizable setting, and thresholds and union of intervals the agnostic setting. We also examine a different formulation of comparing SSL with supervised learning with negative conclusions for SSL.

Chapter 5. We finally conclude by taking a step back and providing general commentary on the big picture of semi-supervised learning and what insights our results give. We describe three open questions for future research into the theory of semi-supervised learning, and offer some general directions researchers can take.

[2] when no assumptions are made on the relationship between labels and the unlabeled data structure.

Chapter 2

Background in Statistical Learning Theory

Before presenting a new framework for the theoretical analysis of semi-supervised learning, we will first review some background material on supervised learning theory, also known as *statistical learning theory*. Our framework will be an extension of the usual *Probably Approximately Correct* (PAC) model of Valiant (1984) for supervised learning. We will also cover its "agnostic" version (Haussler, 1992; Kearns et al., 1992, 1994). Since this thesis is mostly concerned with the informational or statistical aspects of learning—either supervised or semi-supervised—we avoid issues of computational complexity of learning, the study of such issues along with statistical aspects is sometimes known as *computational learning theory*.

For a more comprehensive and pedagogical treatment of the material in this chapter, we refer the reader to the following expository works: Anthony and Bartlett (1999); Kearns and Vazirani (1994); Devroye et al. (1997); Vapnik (1998).

In the rest of this chapter, we will first lay down the notation in Section 2.1 to be used throughout the thesis. Then in Section 2.2 we give an expository tour of the motivation behind the various aspects and definitions of the PAC model. In Section 2.3 and 2.4 we formally define notions of a *learning algorithm* and the *sample size requirements* of learning algorithms. Finally, in Section 2.5 we describe the incredible connection between characterizing learning and the seminal work of Vapnik and Chervonenkis (1971) on the foundations of statistics.

2.1 Some Notation

While we will also develop notation for new definitions found throughout the remainder of this thesis, we will now present some notation that can be digested for the reader without a background in statistical learning theory. For background in probability theory at the measure theoretic level, see for example (Billingsley, 1995).

- Let $\mathcal{X}$ denote the input domain of a classification problem, $\mathcal{Y} = \{0, 1\}$, and $\mathfrak{S}_0$ a σ-algebra over $\mathcal{X}$. We define a measure space $(\mathcal{X} \times \mathcal{Y}, \mathfrak{S})$ where the σ-algebra $\mathfrak{S}$ consists of sets of the form $A \times B$ where $A \in \mathfrak{S}$ and $B \subseteq \mathcal{Y}$. Typically, we will assume the input data to a learning algorithm are drawn independently from some probability measure over this space (see Section 2.2 for more details).

- For a probability distribution P, we denote by P^m the product distribution $\underbrace{P \times \cdots \times P}_{m \text{ times}}$.

- For a random variable X,
  - we denote $X \sim P$ if X is distributed according to a probability distribution P,
  - we denote $\Pr_{X \sim P}(A)$ the probability that $X \in A$ of a measurable set A,
  - and we denote the expectation of X with respect to P by $\mathbb{E}_{X \sim P}(X)$.

- For a probability distribution P over $X \times Y$ we denote $P(Y|X)$ the conditional distribution of Y given X.

- For a positive integer n, denote $[n] = \{1, 2, \ldots, n\}$.

- We use $:=$ to indicate a definition of an equality.

- For a subset T of a domain set, we use $\mathbf{1}T$ to denote its characteristic function (i.e. equals to 1 if $x \in T$ and 0 otherwise).

- We use $\mathbb{R}$ and $\mathbb{N}$ to denote the real numbers and non-negative integers, respectively.

- We use $O(\cdot)$ for the big-O notation, $\Omega(\cdot)$ for big-omega notation and $\Theta(\cdot)$ for big-theta notation.

- We use $\circ$ for function composition.

- For an indicator function I over some domain $\mathcal{X}$, we define $\mathsf{set}(I) := \{x \in \mathcal{X} : I(x) = 1\}$.

- For two sets A and B we denote their symmetric difference by $A \Delta B = (A \backslash B) \cup (B \backslash A)$.

The next section we will discuss some issues in formalizing a model of computational learning, and hence motivate the concepts of the Probably Approximately Correct model.

2.2 Motivating the Probably Approximately Correct Model

Let us begin with some basic definitions.

Definition 2.1. The *labeled training data* is a collection

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

where for each $i \in [n]$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Each element of the collection is known as an *example*.

In this thesis, we will let $\mathcal{Y} = \{0, 1\}$, which is known as the *classification* setting. When $\mathcal{Y}$ is a continuous set, then it is known as the *regression* setting.

For convenience we will sometimes refer to labeled training data as *labeled data*, *labeled sample* or simply *training data*. As an example of the above definitions, if we are interested in automatically classifying email as spam or not spam, we can model this by letting $\mathcal{X}$ be the set of all possible emails (e.g. represent it as a bag of words vector), 0 as not spam and 1 as spam. We want to design a spam detecting algorithm that takes as input labeled training data and outputs a predictor for future emails.

Definition 2.2. A *hypothesis* (or *predictor*, or *classifier*) is a function $h : \mathcal{X} \to \mathcal{Y}$ such that $h^{-1}(1) \in \mathfrak{S}_0$, for a σ-algebra $\mathfrak{S}_0$ over $\mathcal{X}$. That is, the *set representation of* h, denoted by $\mathsf{set}(h) = h^{-1}(1)$, is measurable. A *hypothesis class* (or *space*) [1] $\mathcal{H}$ is a set of hypotheses.

Intuitively the input training data must be "representative" of future emails otherwise you cannot learn useful rule for future prediction. For example, one cannot expect a student to do well in an exam on linear algebra if a classroom teacher always gives homework questions on calculus. The PAC model overcomes this issue by assuming that the training data and future test data are sampled *i.i.d.* (independently, identically distributed) according to some fixed, but unknown (to the learning algorithm) distribution P.

Definition 2.3. A *data generating probability distribution*, or a *labeled distribution*, P is a probability measure defined over $(\mathcal{X} \times \mathcal{Y}, \mathfrak{S})$.

Thus, the input training data is a random variable distributed according to P^m where m is the training data size. Our aim is to design a learning algorithm that given the training data, outputs a hypothesis h with low future error on examples from P.

[1] Technically we should require that $\mathcal{H}$ be *permissible*, a notion introduced by Shai Ben-David in (Blumer et al., 1989) which is a "weak measure-theoretic condition satisfied by almost all real-world hypothesis classes" that is required for learning.

Definition 2.4. Let P be a distribution over $\mathcal{X} \times \mathcal{Y}$, and S be a labeled training sample. The *true error of a hypothesis* h with respect to P is

$$\mathrm{Er}^P(h) := \mathop{\mathbb{E}}_{(x,y)\sim P}(h(x) \neq y) = P\{(x,y) : h(x) \neq y\}.$$

The *empirical error of a hypothesis* h with respect to S is the fraction of misclassified examples,

$$\mathrm{Er}^S(h) := \frac{|\{(x,y) \in S : h(x) \neq y\}|}{|S|}.$$

Note that the curly braces above represent a collection rather than a set. The *Bayesian optimal* hypothesis with respect to P is

$$\mathrm{OPT}_P(x) = \begin{cases} 0 & \text{if } P(y=0|x) \geq 1/2, \\ 1 & \text{otherwise} \end{cases}.$$

It is easy to see that the Bayesian optimal classifier is the function with the smallest possible error. Of course, we usually do not have access to P otherwise the learning problem simply becomes outputting the Bayesian optimal.

Now we are almost ready to define what it is for an algorithm to learn. One attempt is to say that an algorithm learns if when we are given more and more labeled data, we can output hypotheses that get closer and closer in true error to the Bayesian optimal. However, there is one huge problem: a phenomenon known as *overfitting*.

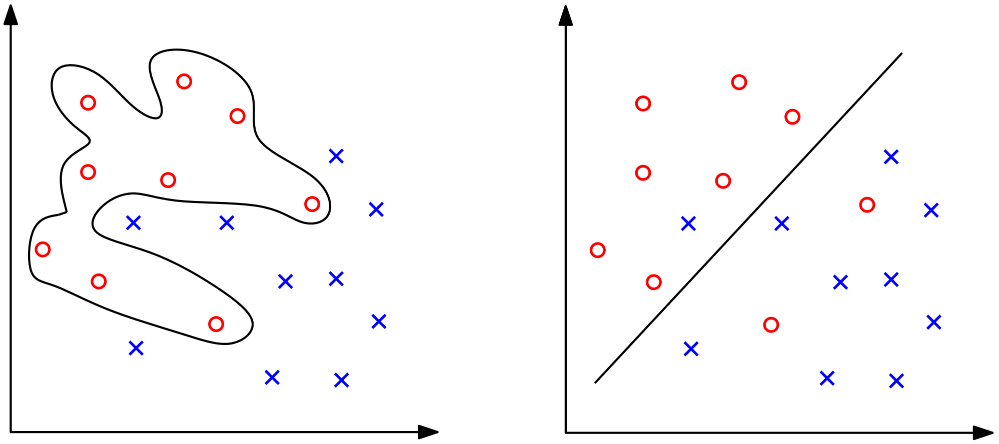

Figure 2.1: Which predictor is better—the "complicated" shape which makes no mistakes on separating the training data or the simple linear separator that makes small mistakes? *Occam's razor:* "one should not increase, beyond what is necessary, the number of entities required to explain anything."

Here is an informal example that captures the essence of overfitting: a classroom teacher gives some questions and answers, there is a student who learns by memorizing all the question and answers, should the student expect to do well on an exam? If the exam questions are the same as those given earlier, then the student will do perfectly, but if the exam has entirely different questions then quite possibly not. The major issue here is that it is difficult for the student to assess her own exam performance if she has not seen the exam, even though she can do perfectly given previously seen questions. In our setting, the exam plays the role of future data to classify while questions and answers are training data. See 2.1 for a pictorial example. Below is a more formal example.

Example 2.1 (Overfitting)**.** Let $\mathcal{X} = [0,1]$ and P be such that its marginal over X is the uniform distribution and for all $x \in [0,1]$, $P(y=1|x) = 1$. For any $m > 0$, suppose a training sample is drawn $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim P^m$. Consider the hypotheses

$$\mathbf{1}\{x_1, \ldots, x_m\} = \begin{cases} 1 & \text{if } \exists i \in [m],\ x = x_i \\ 0 & \text{otherwise} \end{cases}, \qquad\qquad h(x) = 1.$$

Clearly $\mathrm{Er}^P(h) = \mathrm{Er}^S(h) = 0$. On the other hand $\mathrm{Er}^S(\chi_S) = 0$, but $\mathrm{Er}^P(\chi_S) = P\{(x,y) \notin S\} = 1$. However, since the input to any algorithm is only S, it cannot decide how to label the remaining points outside of S because the sample error may be far from the P-error. How does the PAC model fix this?

2.3 The Agnostic PAC Model

The solution to overfitting in the PAC framework is to restrict oneself to a hypothesis class $\mathcal{H}$ that is in some way not “complex” (i.e. cannot be too rich a class that can overfit), and require that a learning algorithm be one in which the larger the input training data, the closer the error of the algorithm's output hypothesis to the *best* hypothesis in $\mathcal{H}$.

Definition 2.5. Let $\mathcal{X}$ be a domain, $\mathcal{Y} = \{0,1\}$, and $\mathcal{H}$ a hypothesis class. A *supervised learning algorithm* with respect to $\mathcal{H}$ is an algorithm

$$\mathcal{A} : \bigcup_{i=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^i \to \mathcal{Y}^{\mathcal{X}},$$

($\mathcal{Y}^{\mathcal{X}}$ is the set of all possible functions from $\mathcal{X} \to \mathcal{Y}$) such that for every $(\epsilon, \delta) \in (0,1]^2$, there exists a non-negative integer m such that for all probability distributions P over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over a random sample $S \sim P^m$,

$$\mathrm{Er}^P(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \le \epsilon\ .$$

The parameters ϵ and δ are referred to as the *accuracy* and *confidence* parameters, respectively. The above probability and condition can be succinctly written as

$$\Pr_{S \sim P^m}\left(\mathrm{Er}^P(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \le \epsilon \right) \ge 1 - \delta\ . \tag{2.1}$$

Occasionally we will be using the term *supervised algorithm* to refer to any algorithm that takes training data and outputs a hypothesis (i.e. it does not need to "learn").

Several things to note in the definition. First, the learning requirement is ***distribution-free***: there is a sample size for $\mathcal{A}$ that only depends on $\epsilon, \delta, \mathcal{H}$ such that for *any* data generating distribution P drawing a sample of size $m(\mathcal{A}, \epsilon, \delta, \mathcal{H})$ from P suffices to learn within ϵ accuracy and δ confidence. This may seem like a very strong condition, but one of the crowning achievements of learning theory is that this condition can be satisfied given $\mathcal{H}$ is a "nice" class (see Section 2.5).

Second, the ϵ accuracy condition is ***relative to the "best"*** predictor in $\mathcal{H}$ (the best may not exist, but for any $\xi > 0$ there exist predictors with true errors that are ξ-close). We only require the learner do well with respect to the best in $\mathcal{H}$ and *not necessarily* the Bayesian optimal. Of course, learning would not be interesting if one does not fix a $\mathcal{H}$ whose best hypothesis has small error. In the extreme case, if no hypothesis in $\mathcal{H}$ has less than 50% error, then learning is trivially achieved by outputting a classifier that flips a fair coin to guess a label. Thus,

> *the learner must use her prior knowledge to choose an appropriate $\mathcal{H}$ that contains a hypothesis with small error.*

However, there's another issue, as we have seen in Example 2.1 and as we will see in Section 2.5, if a $\mathcal{H}$ is chosen with "too many" functions in the hopes that it will include a hypothesis with low error, one runs into problems with overfitting. Thus, the learner faces a tradeoff in specifying a $\mathcal{H}$ that contains a hypothesis with low error and not having a class of "rich" functions.

Third, the algorithm can output *any* hypothesis, not just ones inside $\mathcal{H}$. The only requirement is that the true error of the output hypothesis is close to the true error of the "best" hypothesis in $\mathcal{H}$. For example, the algorithm may even restrict its output predictors to lie inside another hypothesis class $\mathcal{H}'$ that does not overlap with $\mathcal{H}$.

While in the original PAC formulation the learning algorithm must run in polynomial time with respect to the training sample size, in this thesis we will ignore any computational complexity issues. Instead we focus on informational complexity.

Definition 2.6. We define the *supervised learning (SL) sample complexity with respect to* supervised algorithm $\mathcal{A}$, $\epsilon, \delta > 0$, hypothesis class $\mathcal{H}$, and distribution P over $\mathcal{X} \times \mathcal{Y}$ as

$$m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P) := \min \left\{ m \;\middle|\; \Pr_{S \sim P^m} \left(\mathrm{Er}^P(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \leq \epsilon \right) \geq 1 - \delta \right\}$$

and we define the *supervised learning sample complexity* as

$$m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) := \min_{\mathcal{A}} \sup_{P} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P).$$

That is, the sample complexity tells us that there exists a learning algorithm such that if the training data is of size at least $m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H})$ then we can obtain accuracy ϵ and confidence δ. If the input data size is any less than $m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H})$ then there is *no algorithm* that can obtain the desired accuracy and confidence. The sample complexity for supervised learning and semi-supervised learning (to be defined later) is the main focus of this thesis. In essence, we are interested in seeing if it is smaller for SSL compared with SL.

2.4 The Realizable PAC Model

What we have described in the previous section is the known as the *agnostic* extension of the original PAC model. It is so-called because one does not assume anything about the Bayesian optimal classifier—it can have positive error or it can have zero error and may not lie inside $\mathcal{H}$. In this section we describe the original PAC model, one that assumes the labels of data generating distribution come from some hypothesis in $\mathcal{H}$.

Definition 2.7. Fix $\mathcal{H}$ over $\mathcal{X}, \mathcal{Y}$. A hypothesis h is a *target hypothesis* if the underlying data generating distribution P over $\mathcal{X} \times \mathcal{Y}$ has the property that $P(y = h(x)|x) = 1$. More succinctly, $\mathrm{Er}^P(h) = 0$. For such a P, we rewrite it as P_h. The *realizable* setting occurs if a *target hypothesis* exists in $\mathcal{H}$.

Indeed, the realizable assumption is quite strong: the learner must know in advance that the Bayesian optimal has zero error and must lie in the chosen hypothesis class. On the other hand, one can obtain better bounds on the sample complexity (see Section 2.5). The modified definition of a learning algorithm is the same as in Definition 2.5 except the set of distributions is restricted to those that respect the realizable property.

Definition 2.8. A *supervised learning algorithm for realizable setting* with respect to $\mathcal{H}$ is an algorithm

$$\mathcal{A} : \bigcup_{i=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^i \to \mathcal{Y}^{\mathcal{X}},$$

such that for every $(\epsilon, \delta) \in (0, 1]^2$, there exists a non-negative integer m such that for all probability distributions P_g over $\mathcal{X} \times \mathcal{Y}$, where $g \in \mathcal{H}$,

$$\Pr_{S \sim P_g^m} \left(\mathrm{Er}^{P_g}(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^{P_g}(h) \le \epsilon \right) \ge 1 - \delta \, .$$

We will often use the term *supervised learning algorithm* to refer to both the realizable and agnostic settings, and it will be clear which definition is being used by the context. The sample complexity can be defined analogously to Definition 2.6 (again, we will use the term SL sample complexity to refer to both agnostic and realizable settings).

Definition 2.9. We define the *supervised learning (SL) sample complexity for realizable setting with respect to* supervised algorithm $\mathcal{A}$, $\epsilon, \delta > 0$, hypothesis class $\mathcal{H}$, and distribution P_g over $\mathcal{X} \times \mathcal{Y}$ where $g \in \mathcal{H}$ as

$$m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P_g) := \min \left\{ m \;\middle|\; \Pr_{S \sim P_g^m} \left(\mathrm{Er}^{P_g}(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^{P_g}(h) \leq \epsilon \right) \geq 1 - \delta \right\}$$

and we define the *supervised learning sample complexity for realizable setting* as

$$m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) := \min_{\mathcal{A}} \sup_{P_g : g \in \mathcal{H}} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P_g).$$

Going back to address the issue in Example 2.1 on overfitting, we will now turn our attention to "complexity" or "richness" property of $\mathcal{H}$ that will affect both the issue of whether a learning algorithm exists at all and how the sample complexity depends on it.

2.5 Learnability and Distribution-Free Uniform Convergence

It turns out that there is a beautiful connection between PAC learning and the seminal work of Vapnik and Chervonenkis (1971) on empirical process theory. Roughly, their result says that given $\mathcal{H}$ is not too complex (to be defined shortly) then it is possible to estimate the true error of *all* hypotheses via the sample error given a sufficiently large sample that *does not* depend on the data generating distribution. Before we formally state this main theorem, we need to develop the notion of the complexity of $\mathcal{H}$ that will characterize PAC learning.

Definition 2.10. Let $\mathcal{X}$ be a domain, $\mathcal{Y} = \{0, 1\}$, and $\mathcal{H}$ a hypothesis class over $\mathcal{X}, \mathcal{Y}$. For $h \in \mathcal{H}$ and a set $A = \{a_1, \ldots, a_m\} \subseteq X$ let $h(A) := (h(a_1), \ldots, h(a_m)) \in \mathcal{Y}^m$. Define the *growth function* as follows,

$$\Pi(m, \mathcal{H}) := \max_{A \subseteq \mathcal{X} : |A| = m} |\{h(A) : h \in \mathcal{H}\}|.$$

For a set A such that $|\{h(A) \; : \; h \in \mathcal{H}\}| = 2^{|A|}$ we say that $\mathcal{H}$ *shatters* A. The *Vapnik-Chervonenkis dimension* of $\mathcal{H}$ is the size of the largest shatterable set,

$$\mathrm{VC}(\mathcal{H}) := \sup \{m : \Pi(m, \mathcal{H}) = 2^m\}$$

To see some examples of the VC dimension of some basic classes such as union of intervals and linear halfspaces, refer to Appendix A.

The $\mathrm{VC}(\mathcal{H})$ fully characterizes uniform convergence of estimates of sample error to the true error, with explicit upper and lower bounds on the convergence rate that match to within a constant factor. The supremum in the definition of the VC-dimension covers the case when $\mathrm{VC}(\mathcal{H}) = \infty$, in which case, as we will see, imply no learning algorithm exists that can compete with the best hypothesis in $\mathcal{H}$.

2.5.1 Agnostic Setting

Let us first state an upper bound on the uniform convergence. This result was first proved by Vapnik and Chervonenkis (1971) and subsequently improved on with results from Talagrand (1994) and Haussler (1995).

Theorem 2.1 (Distribution-Free Uniform Convergence)**.** *Fix a hypothesis class $\mathcal{H}$ over $\mathcal{X}, \mathcal{Y} = \{0,1\}$. There exists a positive constant C, such that for every $\epsilon, \delta \in (0,1]^2$, if*

$$m \geq \frac{C}{\epsilon^2}\left(\mathrm{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right)$$

then for every probability distribution P over $\mathcal{X} \times \mathcal{Y}$,

$$\Pr_{S \sim P^m}\left(\forall h \in \mathcal{H},\ \left|\mathrm{Er}^S(h) - \mathrm{Er}^P(h)\right| \leq \epsilon\right) \geq 1 - \delta\ . \tag{2.2}$$

One may also express ϵ in terms of given values of m and δ or express δ as a function of ϵ and m. The crucial features of uniform convergence is that it is distribution-*free* as discussed earlier, and the condition inside (2.2) holds for *all* hypotheses in $\mathcal{H}$.

This uniform convergence result leads to a very natural, intuitive, and naive algorithm: given a training sample S, pick any hypothesis h that has the smallest empirical error in $\mathcal{H}$.

Definition 2.11 (ERM Paradigm)**.** Given a training sample S, the *empirical risk minimization* paradigm is any algorithm that outputs a hypothesis with minimum sample error,

$$\mathrm{Er}^S(\mathsf{ERM}(S)) = \min_{h \in \mathcal{H}} \mathrm{Er}^S(h)\ .$$

This is only a statistical principle and does not consider the computational complexity of finding the empirical minimizer. We note that ERM actually refers to a class of possible algorithms that outputs the empirical error minimizer as our next example shows.

Example 2.2 (Different ERM Algorithms)**.** Suppose we are learning *initial segments* (or *thresholds*) over $\mathcal{X} = \mathbb{R}$, that is, hypotheses

$$H = \{\mathbf{1}(\infty, a]\ :\ a \in \mathbb{R}\}$$

and our data generating distribution P is such that the marginal P over $\mathbb{R}$ is the uniform distribution over the unit interval, and the conditional

$$P(y = 1|x) = \begin{cases} 1 & \text{if } x < 1/2 \\ 0 & \text{otherwise} \end{cases},$$

so that $\mathbf{1}(-\infty, 1/2]$ has zero error. Now when a training sample

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

is drawn i.i.d. from P, we can have several different ERM approaches. For simplicity reorder the indices of the sample so that $x_1 \leq x_2 \leq \cdots \leq x_m$, and let $\ell = \max\{x_i : y_i = 1\}$ and $r = \min\{x_i : y_i = 0\}$.

$$\begin{aligned}
\mathsf{LeftERM}(S) &= \mathbf{1}(-\infty, \ell] &(2.3)\\
\mathsf{RightERM}(S) &= \mathbf{1}(-\infty, r) &\\
\mathsf{ERMRandom}(S)(x) &= \begin{cases} 1 & \text{if } x \leq \ell \\ 0 & \text{if } x \geq r \\ 1 & \text{with probability } 1/2 \end{cases} &\\
\mathsf{RandomERM}(S) &\sim \text{Bernoulli}(\{\mathsf{LeftERM}(S), \mathsf{RightERM}(S)\}, 1/2). &(2.4)
\end{aligned}$$

The *deterministic algorithm* $\mathsf{ERMRandom}$ outputs a stochastic classifier that outputs a random guess in the interval $[x_\ell, x_r]$, this is different than the *randomized algorithm* $\mathsf{RandomERM}$ that flips a fair coin and outputs the classifier $\mathsf{LeftERM}(S)$ or $\mathsf{RightERM}(S)$. Note that we need not output a hypothesis in $\mathcal{H}$, just one that has smallest empirical error with respect to those in $\mathcal{H}$ (of course we need to be careful with overfitting). For the rest of this thesis, we will assume that ERM chooses a hypothesis in $\mathcal{H}$ with smallest empirical error.

Now, this is the stage where we show the connection between uniform convergence and sample complexity. Namely, that uniform convergence justifies the ERM principle, which in turn imply an upper bound on the SL sample complexity.

Theorem 2.2 (Agnostic Supervised Learning Upper Bound)**.** *Let C be the constant in Theorem 2.1, and*

$$m_0 := \frac{4C}{\epsilon^2}\left(\mathrm{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right) . \tag{2.5}$$

If $m \geq m_0$ then for any distribution P over $\mathcal{X} \times \mathcal{Y}$,

$$\Pr_{S \sim P^m}\left(\mathrm{Er}^P(\mathsf{ERM}(S)) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \leq \epsilon\right) \geq 1 - \delta .$$

In other words the SL sample complexity $m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) \leq m_0$.

Proof. Let[2] $h_P^* := \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{Er}^P(h)$. By Theorem 2.1, assuming $m \geq m_0$, then for all P, with probability $1 - \delta$ over sample S, the following holds,

$$\begin{aligned}
\mathrm{Er}^P(\mathsf{ERM}(S)) &\leq \mathrm{Er}^S(\mathsf{ERM}(S)) + \frac{\epsilon}{2} && \text{by (2.2)}\\
&\leq \mathrm{Er}^S(h_P^*) + \frac{\epsilon}{2} && \text{by definition of } \mathsf{ERM}\\
&\leq \left(\mathrm{Er}^P(h_P^*) + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} && \text{by (2.2)}\\
&= \mathrm{Er}^P(h_P^*) + \epsilon ,
\end{aligned}$$

which is exactly the requirement from (2.1) for learning. □

[2] While the optimal hypothesis in $\mathcal{H}$ with respect to P may not exist, we can take a hypothesis that gets close enough to the infimum for the proof to follow through.

It turns out that ERM is optimal with respect to the SL sample complexity, up to constant factors. The following is a corresponding lower bound for the sample complexity (i.e. no algorithm has better than constant sample complexity). This lower bound essential results from Vapnik and Chervonenkis (1974).

Theorem 2.3 (Agnostic Supervised Learning Lower Bound)**.** *For $(\epsilon, \delta) \in (0, 1/64)^2$,*

$$m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) \geq \frac{\mathrm{VC}(\mathcal{H})}{320\epsilon^2}$$

if $\mathcal{H}$ contains at least two hypotheses then for $\epsilon \in (0,1)$ and $\delta \in (0, 1/4)$, we also have

$$m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) \geq 2 \left\lfloor \frac{1-\epsilon^2}{\epsilon^2} \ln \frac{1}{8\delta(1-2\delta)} \right\rfloor$$

The proofs of Theorem 2.1 and Theorem 2.3 can be found in the expository book of Anthony and Bartlett (1999, Chap. 4 and 5, respectively). Now we will turn our focus to the sample complexity of the *realizable* setting where one makes the strong assumption that a zero error hypothesis lies in the chosen hypothesis space.

2.5.2 Realizable Setting

For the realizable setting (see the definition in Section 2.4) there are better sample complexity bounds. Basically, the ϵ^2 in the denominator of the sample complexity upper bound (see Equation (2.5)) reduces to ϵ. The corresponding algorithm which matches the sample complexity upper bounds is, unsurprisingly, ERM. The following upper bound was originally proved by Blumer et al. (1989).

Theorem 2.4 (Realizable Supervised Learning Upper Bound)**.** *The supervised learning sample complexity, $m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H})$, for the realizable setting satisfies*

$$m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H}) \leq \frac{4}{\epsilon}\left(\mathrm{VC}(\mathcal{H}) \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right) .$$

This bound is achieved by ERM *which outputs any hypothesis that is consistent with the training sample.*

The following lower bound, which matches the upper bound up to a factor of $\ln(1/\epsilon)$, first appeared in Ehrenfeucht et al. (1989).

Theorem 2.5 (Realizable Supervised Learning Lower Bound)**.** *The supervised learning sample complexity, $m_{\mathrm{SL}}(\epsilon, \delta, \mathcal{H})$, for the realizable setting satisfies*

$$m_{\mathrm{SL}}(\epsilon, \delta) \geq \frac{\mathrm{VC}(\mathcal{H}) - 1}{32\epsilon}$$

and if $\mathcal{H}$ contains at least two hypotheses then for $\epsilon \in (0, 3/4)$ and $\delta \in (0, 1/100)$,

$$m_{\mathrm{SL}}(\epsilon, \delta) > \frac{1}{2\epsilon} \ln\left(\frac{1}{\delta}\right) .$$

The proofs of these bounds can also be found in (Anthony and Bartlett, 1999, Chap. 4 and 5).

Now we are ready to move onto the next chapter of this thesis, where we will present a new model for semi-supervised learning that is based on the PAC model discussed in this chapter.

Chapter 3

Modelling Semi-Supervised Learning

In this chapter we will present a new, formal mathematical model for semi-supervised learning. This model is "utopian," where we assume that the distribution of the unlabeled data is given to the semi-supervised learner. The intuition here is that unlabeled data is abundant and cheap and that the learner can essentially "reconstruct" the distribution over the unlabeled data. From this point of view of semi-supervised learning, any positive result in the usual model of SSL, where a sample of unlabeled data is given rather than an entire distribution, is a positive result in our utopian model. And of course, any inherent limitations of this utopian SSL model implies inherent limitations in the usual model. Our model is based on the Probably Approximately Correct (PAC) model of Valiant (1984) and also its agnostic version (Haussler, 1992; Kearns et al., 1992, 1994). See Chapter 2 for background on PAC learning.

The purpose of this new model is to analyze SSL in a clean way without dealing with unlabeled data sampling issues, and to also compare the potential gains of this utopian model of SSL with that of supervised learning. This last part will be investigated further in Chapter 4. This chapter will be devoted to presenting the model, discussions about the model, and how it fits with related work on the practice and theory of SSL.

In Section 3.1 we present the new model of semi-supervised learning, its motivations and consequences for the ordinary model of semi-supervised learning. We discuss related work in Section 3.2 on the theory of semi-supervised learning and how they are unsatisfactory, as well as putting into perspective various approaches—both practically and theoretically inspired—in the use of unlabeled data. In Section 3.3 we investigate (naturally occurring) scenarios that can hurt semi-supervised learning under the practically popular "cluster assumption," and show the inherent dangers with using inaccurate prior knowledge about the relationship between labels and the unlabeled data structure.

3.1 Utopian Model of Semi-Supervised Learning

In the *usual* setup for semi-supervised learning a learner is given training data that consists of labeled data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$$

and in addition, *unlabeled* data

$$x_{m+1}, x_{m+2}, \ldots, x_{m+u} \in \mathcal{X} \ ,$$

and the goal, or rather hope, is to output a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that is better than what we can get from the labeled data alone. In applications such as natural language processing or computer vision the unlabeled data is typically much more abundant than labeled data. Whereas labeled data may need to be obtained through the use of specialized human annotators (e.g. labeling web pages), unlabeled data is typically widely available (e.g. web pages, emails). Let us first define some terms before continuing our discussion.

Definition 3.1. An *unlabeled probability distribution* D is a distribution over $\mathcal{X}$. For a labeled distribution P over $\mathcal{X} \times \mathcal{Y}$, we denote $\mathcal{D}(P)$ the *marginal distribution* of P over $\mathcal{X}$. That is, for every measurable set $A \subseteq X$, $\mathcal{D}(P)(A) := P(A \times \mathcal{Y})$. The *extension* of an unlabeled distribution D is

$$\mathsf{Ext}(D) := \{P : P \text{ is a distribution over } \mathcal{X} \times \mathcal{Y},\ \mathcal{D}(P) = D\}.$$

For an unlabeled distribution D and a hypothesis h let D_h be the distribution over $\mathcal{X} \times \mathcal{Y}$ such that $D_h(y = h(x) \mid x) = 1$. For an unlabeled sample $S = \{x_1, \ldots, x_m\}$, we denote by $(S, h(S))$ the labeled sample $\{(x_1, h(x_1)), \ldots, (x_m, h(x_m))\}$.

Of course, to provide theoretical guarantees for semi-supervised learning one must make restrictions on the data generating process just as in the PAC model. A very natural extension is to assume that there is an underlying data generating distribution P from which the labeled training data is drawn i.i.d. and also the underlying unlabeled distribution $\mathcal{D}(P)$ from which the unlabeled data is drawn i.i.d.

Because unlabeled data is usually plentiful, we are going to make the "utopian" assumption that the *unlabeled distribution is fully given to the learner.* That is, the learner gets as input a sample of labeled data from P and also the complete distribution $\mathcal{D}(P)$. While this may seem like providing the learner with too much help, we show that there is still limitations of what the learner can gain over not knowing the unlabeled distribution at all. We are being somewhat informal with what it means to compute with distributions (whose support can be continuous), but basically the algorithm has access to the probability of any measurable set as well as samples drawn i.i.d.

One can imagine that in this model, we can obtain so much unlabeled data that we can "reconstruct" the distribution. Of course, this is not completely true because we cannot uniformly estimate the probability of *all* measurable sets (i.e. there is always overfitting) with finite sample. But it gives a framework in which:

1. We steer clear of the sampling issues
2. Any negative result in our model implies negative results in the usual semi-supervised learning model
3. Any positive result in the usual model implies a positive result in our model.

Definition 3.2. Let $\mathcal{H}$ be a hypothesis class, and $\mathfrak{D}$ the set of all unlabeled distributions over $\mathcal{X}$. A *semi-supervised (SSL) learning algorithm* is an algorithm

$$\mathcal{A} : \bigcup_{i=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^i \times \mathfrak{D} \to \mathcal{Y}^{\mathcal{X}},$$

such that for every $(\epsilon, \delta) \in (0,1]^2$, there exists a non-negative integer m such that for all probability distributions P over $\mathcal{X} \times \mathcal{Y}$,

$$\Pr_{S \sim P^m} \left(\mathrm{Er}^P(\mathcal{A}(S, \mathcal{D}(P))) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \le \epsilon \right) \ge 1 - \delta \, .$$

We will also use term *semi-supervised algorithm* to refer to an algorithm that take as input a labeled sample and an unlabeled distribution and outputs a hypothesis (but need not "learn").

In theory, the definition allows the semi-supervised algorithm to output hypotheses not belonging to $\mathcal{H}$. In fact, the algorithm can construct a hypothesis class $\mathcal{H}_D$ dependent on unlabeled distribution D, and output a hypothesis in $\mathcal{H}_D$, but as will be defined later on, the performance must be measured with respect to a hypothesis class fixed before seeing D. This requirement is necessary for a well-defined, and meaningful comparison between supervised and semi-supervised learning (see Section 4.1).

A semi-supervised learning algorithm for the realizable setting (see Section 2.4) can also be analogously defined, except the requirement "for all probability distributions P" it can be replaced with "for all probability distributions D_h where $h \in \mathcal{H}$."

Definition 3.3. We define the *semi-supervised learning (SSL) sample complexity with respect to* semi-supervised algorithm $\mathcal{A}$, $\epsilon, \delta > 0$, hypothesis class $\mathcal{H}$, and distribution P over $\mathcal{X} \times \mathcal{Y}$ as

$$m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P) := \min \left\{ m \;\middle|\; \Pr_{S \sim P^m} \left(\mathrm{Er}^P(\mathcal{A}(S, \mathcal{D}(P))) - \inf_{h \in \mathcal{H}} \mathrm{Er}^P(h) \le \epsilon \right) \ge 1 - \delta \right\}$$

and we define *semi-supervised learning (SSL) sample complexity* as

$$m_{\mathrm{SSL}}(\epsilon, \delta, \mathcal{H}) := \min_{\mathcal{A}} \sup_{P} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P).$$

Remark. We note that this sample complexity is only concerned with the amount of *labeled* data needed, since we are not concerned about the supply of unlabeled data. We can also extend this definition to the realizable setting by restricting the set of data generating distributions to those P where there is some $h \in \mathcal{H}$ with $\mathrm{Er}^P(h) = 0$.

It turns out that the semi-supervised learning sample complexity $m_{\mathrm{SSL}}(\epsilon, \delta, \mathcal{H})$ (on *worst* P) is uninteresting as it can be shown that the worst distribution gives sample complexity bounds as bad as that of supervised learning. Specifically, the bad distribution is concentrated at shattered points with very high noise. See Corollary 4.10 for the sample complexity. Instead we are interested in comparing the sample complexity of SL and SSL for *any fixed* unlabeled distribution. Before moving further to address this question, we discuss some previous semi-supervised learning paradigms—how they are related to our model and how they differ.

3.2 Related Work

Analysis of performance guarantees for semi-supervised learning can be carried out in two main setups. The first focuses on the unlabeled data distribution and does not make any prior assumptions about the conditional label distribution. The second approach focuses on assumptions about how the conditional labeled distribution is related to the unlabeled distribution, under which semi-supervised learning has potentially better label prediction performance than learning based on just labeled samples. The investigation of the first setup was pioneered by Vapnik in the late 1970s in his model of *transductive learning* (see for example, Chapelle et al., 2006, Chap. 24). There has been growing interest in this model in the recent years due to the popularity of using unlabeled data in practical label prediction tasks. This model assumes that unlabeled examples are drawn i.i.d. from an unknown distribution, and then the labels of some randomly picked subset of these examples are revealed to the learner. The goal of the learner is to label the remaining unlabeled examples minimizing the error. The main difference between this model and SSL is that the error of learner's hypothesis is judged only with respect to the known initial sample and not over the entire input domain $\mathcal{X}$.

However, there are no known bounds in the transductive setting that are strictly better than supervised learning bounds[1]. The bounds in Vapnik (2006) are almost identical. El-Yaniv and Pechyony (2007) prove bounds that are similar to the usual

[1] We are not formal about this, but we are comparing the accuracy parameter ϵ (between training error and test data error) in the uniform convergence bounds of transductive learning with that of supervised learning.

margin bounds using Rademacher complexity, except that the learner is allowed to decide *a posteriori* the hypothesis class given the unlabeled examples. But they do not show whether it can be advantageous to choose the class in this way. Their earlier paper (El-Yaniv and Pechyony, 2006) gave bounds in terms of a notion of *uniform stability* of the learning algorithm, and in the broader setting where examples are not assumed to come i.i.d. from an unknown distribution. But again, it's not clear whether and when the resulting bounds are better than the supervised learning bounds.

Kääriäinen (2005) proposes a method for semi-supervised learning, in the realizable setting, without prior assumption on the conditional label distributions. The algorithm of Kääriäinen is based on the observation that one can output the hypothesis that minimizes the unlabeled distribution probabilities of the symmetric difference to all other hypothesis of the *version space* (*i.e.* consistent hypothesis). This algorithm *can* reduce the error of empirical risk minimization by a factor of two. For more details on this algorithm, see Definition 4.5 and the discussion that follows.

Earlier, Benedek and Itai (1991) presented a model of "learning over a fixed distribution." This is closely related to our model of semi-supervised learning, since once the unlabeled data distribution is fixed, it can be viewed as being known to the learner. The idea of Benedek and Itai's algorithm is to construct a minimum ϵ-cover of the hypothesis space under the pseudo-metric induced by the data distribution. The learning algorithm they propose is to apply empirical risk minimization on the hypotheses in such a cover. Of course this ϵ-cover algorithm requires knowledge of the unlabeled distribution, without which the algorithm reduces to ERM over the original hypothesis class. We will explain this algorithm in more detail below (See Section 3.2.1).

The second, certainly more popular, set of semi-supervised approaches focuses on assumptions about the conditional labeled distributions. A recent extension of the PAC model for semi-supervised learning proposed by Balcan and Blum (2005, 2006) attempts to formally capture such assumptions. They propose a notion of a compatibility function that assigns a higher score to classifiers which "fit nicely" with respect to the unlabeled distribution. The rationale is that by narrowing down the set of classifiers to only compatible ones, the complexity of the set of potential classifiers goes down and the generalization bounds of empirical risk minimization over this new hypothesis class improve. However, since the set of potential classifiers is trimmed down by a compatibility threshold, if the presumed label-structure relationship fails to hold, the learner may be left with only poorly performing classifiers. One serious concern about this approach is that it provides no way of verifying these crucial modelling assumptions. In Section 3.3 we demonstrate that this approach may damage learning even when the underlying assumptions seem to hold. In Lemma 3.1 we show that without prior knowledge of such relationship that the Balcan and Blum approach has poor worst-case generalization performance.

Other investigations into theoretical guarantees of semi-supervised learning have shed some light on specific approaches and algorithms. Typically, these generalization bounds become useful when some underlying assumption on the structure of the labeled distribution is satisfied. In the rest of the discussion we will point out some of these assumptions used by popular SSL algorithms.

Common assumptions include the *smoothness assumption* and the related *low density assumption* (Chapelle et al., 2006) which suggests that a good decision boundary should lie in a low density region. For example, Transductive Support Vector Machines (TSVMs) proposed by Joachims (1999) implements the low density assumption by maximizing the margin with respect to both the labeled and unlabeled data. It has, however, been observed in experimental studies (T. Zhang, 2000) that performance degradation is possible with TSVMs.

In Section 3.3, we give examples of mixtures of two Gaussians showing that the low density assumption may be misleading even under favourable data generation models, resulting in low density boundary SSL classifiers with larger error than the outcome of straightforward supervised learning that ignores the unlabeled data. *Such embarrassing situations can occur for any method implementing this assumption.*

Co-training Blum and Mitchell (1998) is another empirically successful technique which has been backed by some generalization bounds by Dasgupta et al. (2001). The idea is to decompose an example x into two "views" (x_1, x_2) and try to learn two classifiers that mostly agree in their classification with respect to the two views. The crucial assumption on the labeled distribution is that x_1 occurs independently from x_2 given the class labels.

Cozman and Cohen (2006) investigate the risks of using unlabeled data in explicitly fitting parametric models (e.g. mixture of Gaussians) $P(x, y \mid \theta)$, where θ is some parameter, to the ground truth data generating distribution $P(x, y)$ (*i.e.* learning generative classifiers). They show a result of the form: if $P(x, y)$ does belong in the parametric family, then the use of unlabeled data will indeed help, otherwise using unlabeled data can be worse than simply ignoring it and performing supervised learning. The former (positive) statement has been known since the late 1970s and is also shown in Castelli (1994); Castelli and Cover (1996); Ratsaby and Venkatesh (1995). Their result attempts to explain experimental observations that show unlabeled data can degrade the performance of generative classifiers (e.g. Bayes nets) when the wrong modelling assumptions are made (see for example, Bruce, 2001).

However, all these approaches, are based on very strong assumptions about the relationship between labels and the unlabeled distribution. These are assumptions that are hard to verify, or to justify on the basis of prior knowledge of a realistic learner.

3.2.1 Previous Theoretical Approaches

Previous approaches to semi-supervised learning for the case when no assumptions are made on the relation between labels and the unlabeled data structure (which we call the *no-prior-knowledge* setting) have used the unlabeled sample to figure out the “geometry” of the hypothesis space with respect to the unlabeled distribution. A common approach is to use that knowledge to reduce the hypothesis search space. In doing so, one may improve the generalization upper bounds.

Definition 3.4. Given an unlabeled distribution D and a hypothesis class $\mathcal{H}$ over some domain $\mathcal{X}$, an *ϵ-cover* is a subset $\mathcal{H}' \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$ there exists $g \in \mathcal{H}'$ such that

$$D(\mathsf{set}(g) \Delta \, \mathsf{set}(h)) \leq \epsilon.$$

Note that if $\mathcal{H}'$ is an ϵ-cover for $\mathcal{H}$ with respect to D, then for every extension $P \in \mathsf{Ext}(D)$,

$$\inf_{g \in H'} \mathrm{Er}^P(g) \leq \inf_{h \in H} \mathrm{Er}^P(h) + \epsilon.$$

In some cases the construction of a small ϵ-cover is a major use of unlabeled data. Benedek and Itai (1991) analyze the approach, in the case when the unlabeled distribution is fixed and therefore can thought of as being known to the learner. They show that the smaller an ϵ-cover is, the better its generalization bound for the empirical risk minimization (ERM) algorithm over this cover. However, is it the case that if the ϵ-cover is smaller, then supervised ERM can also do just as good? For example, if the unlabeled distribution has support on one point, the ϵ-cover will have size at most two, and while the sample complexity goes down for the fixed distribution learner, it also goes down for an oblivious supervised ERM algorithm. This latter question is what differentiates the focus of this thesis from the focus of fixed distribution learning.

Balcan and Blum (2006) suggest a different way of using the unlabeled data to reduce the hypothesis space. However, we claim that without making any prior assumptions about the relationship between the labeled and unlabeled distributions, their approach boils down to the ϵ-cover construction described above.

Lemma 3.1. *Let $\mathcal{H}$ be any hypotheses class, $\epsilon, \delta > 0$, and D be any unlabeled distribution. Let $\mathcal{H}' \subseteq \mathcal{H}$ be the set of “compatible hypotheses.” Suppose $\mathcal{A}$ is an SSL algorithm that outputs any hypothesis in $\mathcal{H}'$. If $\mathcal{H}'$ does not contain an ϵ-cover of $\mathcal{H}$ with respect to D, the error of the hypothesis that $\mathcal{A}$ outputs is at least ϵ regardless of the size of the labeled sample.*

Proof. Since $\mathcal{H}'$ does not contain an ϵ-cover of $\mathcal{H}$, there exist a hypothesis $h \in H$ such that for all $g \in H'$, $D(\mathsf{set}(g) \Delta \, \mathsf{set}(h)) > \epsilon$. Thus, for any $g \in H'$, $\mathrm{Er}^{D_h}(g) > \epsilon$. Algorithm $\mathcal{A}$ outputs some $g \in H'$ and the proof follows. □

This lemma essentially says that either the "compatible hypotheses" form an ϵ-cover—in which case it reduces to the algorithm of Benedek and Itai (1991) that is not known to do better than supervised learning, or it hurts SSL by learning over a hypothesis class $\mathcal{H}'$ whose best hypothesis is ϵ worse than the best in $\mathcal{H}$.

Kääriäinen (2005) utilizes the unlabeled data in a different way. Given the labeled data his algorithm constructs the version space $V \subseteq \mathcal{H}$ of all sample-consistent hypotheses, and then applies the knowledge of the unlabeled distribution D to find the "centre" of that version space. Namely, a hypothesis $g \in V$ that minimizes $\max_{h \in \mathcal{H}} D(\mathsf{set}(g) \Delta \mathsf{set}(h))$. See Definition 4.5 and the discussion that follows for a concrete example of this algorithm over thresholds on the real line.

Clearly, all the above paradigms depend on the knowledge of the unlabeled distribution D. In return, better upper bounds on the sample complexity of the respective algorithms (or equivalently on the errors of the hypotheses produced by such algorithms) can be shown. For example, Benedek and Itai (1991) give (for the realizable case) an upper bound on the sample complexity that depends on the size of the ϵ-cover—the smaller ϵ-cover, the smaller the upper bound.

In the next section we analyze a concrete example of the issues inherent in doing semi-supervised learning with assumptions on the relationship between labels and the unlabeled distribution. This assumption, known as the *cluster assumption* or sometimes the *low density assumption*, is a widely held belief when performing SSL. However, it may mislead SSL and result in a worse classifier compared to simply performing supervised learning using only the labeled data.

3.3 Issues with the Cluster Assumption

One important point that can be raised against a no-prior-knowledge (i.e. not making assumptions on the relationship between labels and the unlabeled data structure) analysis of semi-supervised learning is that in practice, people often make assumptions on the relationship between labels and the unlabeled distribution. And while it may not be surprising that such a no-prior-knowledge analysis shows that unlabeled can't help by much, it is also worthwhile to note that even when these assumptions do hold to *some* degree, it is still possible for the learner to end up doing much worse than simply ignoring the unlabeled data and performing supervised learning.

One very common assumption is known as the *cluster assumption*, which is loosely defined as the following, taken from Chapelle et al. (2006, Chap.1).

> **Cluster Assumption.** "If points are in the same cluster, they are likely to be of the same class."

This seems like a very natural assumption. Data points that are "clustered" closely together should share the same labels, and two data points that are more distant should belong to different classes (this is not explicitly stated in the assumption, though). Of course, this assumption as stated above depends on a definition of what it means for two points to be in the same "cluster," which is not a well-defined notion in unsupervised learning.

A popular intuition for what clusters are is the belief that if there is a high-density path (with respect to the data generating distribution) between two points, then those two points should belong in the same cluster, otherwise they should belong in different clusters. This gives rise to a similar, sometimes known as being equivalent, SSL assumption (taken from Chapelle et al., 2006, Chap. 1)

> **Low Density Boundary Assumption.** "The decision boundary should lie in a low density region."

For example, the popular Transductive Support Vector Machine of Joachims (1999) uses this assumption. However, this again, is an ill-defined assumption. While the density of a decision boundary is a well-defined mathematical notion, it is not clear what it means for the boundary to lie in *a* low density region. Should it be the lowest density boundary possible or should it have density some ϵ away from the lowest density boundary? What should this tolerance, ϵ, be?

To make the above issues concrete, consider the examples shown in Figures 3.1, 3.2 and 3.3. In Figure 3.1, there is a mixture of two Gaussians on the real line with the same variance but different means. Each Gaussian always generates labels of one type and the other Gaussian generates labels of another type. The Bayesian Optimal separation boundary is at $x = 1$. However, their combined distribution has the highest density at $x = 1$! First, the cluster assumption does not even apply, and second the low density assumption is misleading as the highest density point is the best decision boundary.

Figure 3.2 shows a similar phenomenon, except the two (homogeneous labeled) Gaussians have different variance and mean. Their combined density first reaches a high density point, then drops to a low density point and then rises to another high density point. The cluster assumption and the low density assumption says that there should be two clusters here, separated by the least dense threshold. In this case the lowest density threshold is close to $x = 3$, but the Bayesian Optimal threshold is close to $x = 2$! This results in a significant difference in the error of the low density output and the optimal. Figure 3.3 shows a similar issue, except the mixtures are not Gaussians but still homogeneous "blocks." In this case, the error of the low density classifier is twice as bad as that of the optimal.

In all three of the above examples, a supervised learning algorithm that performs a simple ERM scheme will pick something close to the optimal boundary, given sufficient labeled examples. But semi-supervised learning that implements these assumptions can be mislead and always end up choosing the bad classifier regardless of the size of the labeled examples.

The approach of Balcan and Blum (2005) suffers from the same issue. The threshold of the compatibility function may be such that the remaining compatible

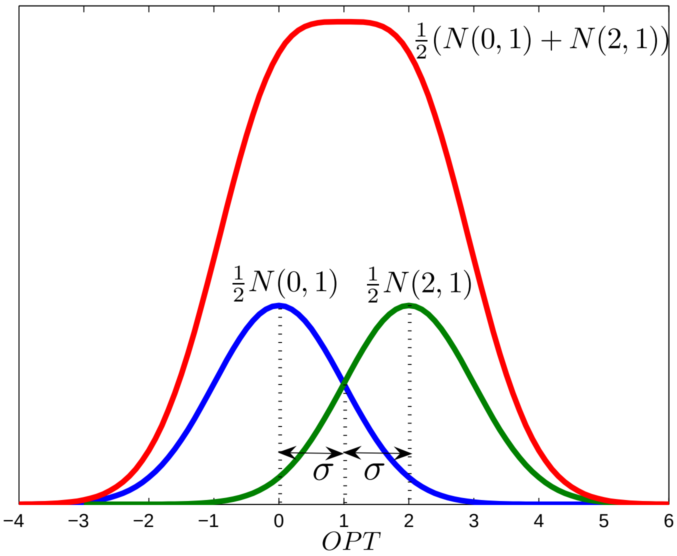

Figure 3.1: Mixture of two Gaussians $\mathcal{N}(0,1)$ (labeled '-') and $\mathcal{N}(2,1)$ (labeled '+') shows that the optimum threshold is at $x = 1$, the densest point of the unlabeled distribution. The sum of these two Gaussians is unimodal.

hypothesess are all bad classifiers. For example, the compatible hypotheses may only be those ones that have the lowest possible density, of which none might be the best classifier.

Rigollet (2007) presents a formal model of the cluster assumption. Given a probability distribution, D over some Euclidean data domain and its corresponding density function f, define, for any positive real number, a, $L(a) = \{x : f(x) > a\}$. The *cluster assumption* says that points in each of the connected components of $L(a)$ (after removal of "thin ribbons") have the same Bayesian optimum label. These components are the "clusters" and the SSL learner Rigollet proposes simply assigns the majority label to each cluster, given the labeled data. This is a very strong assumption under which one uses unlabeled data.

Since Rigollet's SSL learner's hypothesis space is data dependent (*i.e.* all possible labellings of the discovered clusters), it does not fit our framework where the hypothesis space must be fixed before seeing the unlabeled distribution. Thus, it is not really comparable with supervised learning in our setting.

However, in spite of the strong cluster assumption of Rigollet and the data dependent hypothesis space, we can prove that the ratio between the sample complexity of SSL and SL is at most d, the Euclidean dimension of the input data. In particular, the results of Section 4.4 (see Theorem 4.9) show a lower bound of $\Omega\left(\frac{k+\ln(1/\delta)}{\epsilon^2}\right)$ on the sample complexity of SSL learning under this cluster assumption[2], where k is the number of connected components of $L(a)$ (*i.e.* "clusters").

[2] Note that technically this lower bound applies when the unlabeled distribution mass of each

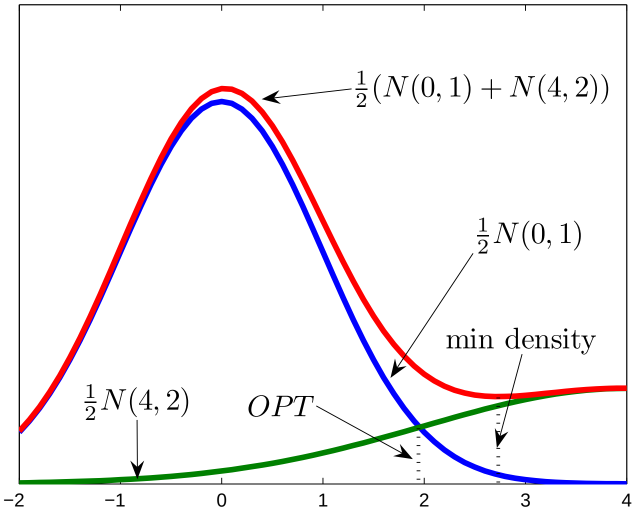

Figure 3.2: Mixture of two Gaussians $\mathcal{N}(0,1)$ (labeled '-') and $\mathcal{N}(4,2)$ (labeled '+') with difference variances. The minimum density point of the unlabeled data (the sum of the two distributions) does not coincide with the optimum label-separating threshold where the two Gaussians intersect. The classification error of optimum is ≈ 0.17 and that of the minimum density partition is ≈ 0.21.

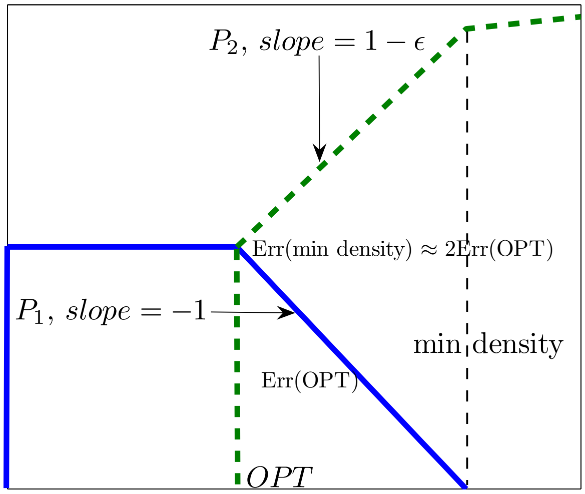
P_2, $slope = 1-\epsilon$
Err(min density) $\approx$ 2Err(OPT)
P_1, $slope = -1$
min density
Err(OPT)
OPT

Figure 3.3: The solid line indicates the distribution P_1 (labeled '-') and the dotted line is P_2 (labeled '+'). The x coordinate of their intersection is the optimum label prediction boundary. The slope of the solid line is slightly steeper than that of the dotted line ($|-1| > 1-\epsilon$). The minimum density point occurs where the density of P_1 reaches 0. The error of the minimum unlabeled density threshold is twice that of the optimum classifier.

For a supervised learner that only has access to labeled examples, the learner can apply a simple ERM algorithm to the class of all k-cell Voronoi partitions of the space. Since the VC-dimension of the class of all k-cell Voronoi partitions in R^d is of order kd, the usual VC-bounds on the sample complexity of such a SL learner is $O\left(\frac{kd+\ln(1/\delta)}{\epsilon^2}\right)$ examples.

In Chapter 4 we will formally analyze what semi-supervised learning can gain over supervised learning (in the labeled sample complexity) under our framework when no assumptions are made between the relationship of labels and the unlabeled distribution (*i.e.* a no-prior-knowledge analysis).

of the clusters are the same, but its proof can be adapted so that there's an extra constant hidden in the lower bound that depends on how "evenly balanced" the cluster probability masses are.

Chapter 4

Inherent Limitations of Semi-Supervised Learning

While it may appear that having complete knowledge of the unlabeled distribution can provide great advantage in the labeled sample complexity required when doing SSL in practice, compared to supervised learning, we give evidence in this chapter that there are in fact inherent limitations of the advantage of having this extra knowledge. There is an important caveat here, we conjecture for any general hypothesis space, and prove it for some basic hypothesis classes, that knowing the unlabeled distribution does not help *without* making assumptions about the relationship between labels and the unlabeled distribution. That is, our analysis in this chapter can be called a *no-prior-knowledge* analysis where one does not make additional assumptions than what is assumed in the typical PAC model (e.g. one does not assume the cluster assumption).

Of course, one may object and say that not making SSL specific assumptions obviously implies unlabeled data is unuseful. However, there are a few points to make about this issue:

1. We prove that the "center of version space" algorithm of Kääriäinen (2005) (that do not make SSL assumptions) for thresholds over the real line can gain a factor of two over currently known upper bounds on the sample complexity of supervised learning.

2. We have shown in Section 3.3 that there is a potential danger of damaging learning when making SSL assumptions. The danger occurs when these assumptions don't fit, if even slightly.

3. The current state of the art in (unlabeled) distribution-specific learning do not provide tight upper and lower bounds on the sample complexity that match, within a constant factor independent of the hypothesis class, the upper bound

of supervised learning, for *any* unlabeled distribution[1]. For example, the lower bound of Benedek and Itai (1991) is very loose.

It is true that when doing SSL in practice, many practitioners make assumptions on the relationship between labels and the unlabeled data distribution (e.g. cluster assumption). But we have seen that this can potentially lead to undesirable situations. It remains to investigate for future research how to decrease the risks of SSL, relax strong assumptions, while guaranteeing significantly better sample complexity. However, this thesis attempts to understand under which scenarios unlabeled data can't help and when it can potentially help.

In Section 4.1 we present an ambitious and fundamental conjecture that proclaims for any fixed unlabeled distribution, the sample complexity of supervised ERM is not much worse than the sample complexity of the best SSL algorithm. This will set the stage for the remainder of the chapter where we prove the conjecture for some basic hypothesis classes over the real line. In Section 4.2 we show that for some natural hypothesis classes over the real line, doing SSL with "nice" unlabeled distributions is equivalent to supervised learning under the uniform distribution. This simplification allows us to prove the conjecture for the realizable case of thresholds in Section 4.3 and also for the agnostic case of thresholds and union of intervals in Section 4.4. Finally we end the chapter in Section 4.5 with an alternate formulation of comparing SSL with SL and show that it still does not provide advantages for SSL.

4.1 Fundamental Conjecture on No-Prior-Knowledge SSL

The conjecture that we propose roughly asserts that semi-supervised learning in our utopian model cannot improve more than a constant factor over the sample complexity of supervised learning on *any* unlabeled distribution. The key point here is that no prior SSL assumptions are held, therefore any noisy labeled distribution is allowed, even ones that for example, do not intuitively satisfy the cluster assumption.

Conjecture 4.1. *In the agnostic setting, there exists constants $C \geq 1$, $\xi, \xi' > 0$, and a fixed empirical risk minimization algorithm ERM_0, such that for all $\mathcal{X}$, $\epsilon \in (0, \xi)$, $\delta \in (0, \xi')$, for every hypothesis class $\mathcal{H}$ over $\mathcal{X}$, for every distribution D over $\mathcal{X}$, we have*

$$\sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SL}}(\mathsf{ERM}_0, \epsilon, \delta, \mathcal{H}, P) \leq C \min_{\mathcal{A}} \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P) \tag{4.1}$$

[1] Some bounds have been proven for specific distributions like the uniform distribution on the unit ball in Euclidean space, however proving tight bounds for any distribution still seems like a far goal.

There are several important things to note about this conjecture.

- **The constant** C. Notice the order of the quantifiers in the statement: the constant C cannot depend on $\mathcal{H}$ (or $\mathcal{X}$, ϵ, δ). If the quantifiers were reversed and allowed to depend on $\mathcal{H}$, and supposing $\mathcal{H}$ contains some h and $\bar{h}$ (i.e. $\bar{h} = 1 - h$) then one can show the conjecture to be true by letting $c = O(\mathrm{VC}(\mathcal{H}))$. This follows from a lower bound of $\Omega(\frac{\ln(1/\delta)}{\epsilon^2})$ for SSL (see Lemma 4.7) and standard distribution-free ERM upper bounds of $O(\frac{\mathrm{VC}(\mathcal{H})+\ln(1/\delta)}{\epsilon^2})$ for supervised learning (see Theorem 2.1).

- **Hypothesis class is same for both SSL and SL**. If we compare the sample complexities of SL under some hypothesis class $\mathcal{H}$ and that of SSL under a hypothesis class *dependent on* D it is not a well-defined or meaningful comparison. For example, if $\mathcal{H}$ consist of a singleton, then SL sample complexity is trivially zero. Meanwhile if the SSL hypothesis class (that is dependent on D) is more complex, it can have much greater sample complexity. Also note that the universal quantifier over the hypothesis class comes before that for the unlabeled distribution.

- **There can be different ERM algorithms**. There can be many hypothesis that have minimum sample error in $\mathcal{H}$, so different ERM algorithms have different rules when choosing. See Example 2.2.

- **Algorithm ERM_0 cannot depend on D, but $\mathcal{A}$ can**. In the conjecture, the order of the quantifiers says that there exists an ERM algorithm ERM_0 before the universal (i.e. for all) quantifier over unlabeled D's. However, the semi-supervised algorithm $\mathcal{A}$ can be dependent on D. In other words, $\mathcal{A}$ can potentially exploit the knowledge of D. And this conjecture says that it can't significantly exploit it to its advantage *for all* unlabeled distributions D, even ones that may behave "nicely."

- **Comparing ERM and best SSL algorithm on *fixed* distribution**. The condition (4.1) must hold *for all* unlabeled distributions. Given *any* fixed D, the condition is essentially saying that the worst conditional distribution over D, with respect to sample complexity, for supervised ERM_0 is *not much worse* than the worst conditional distribution for the best SSL algorithm for D.

- **ERM sample complexity may be better than** $O(\mathrm{VC}(\mathcal{H})/\epsilon^2)$. For an unlabeled distribution with support on a single point, the best algorithm for both SL and SSL is to predict the majority. Thus, the two algorithms are the same and have the same sample complexity of $\Theta(\frac{\ln(1/\delta)}{\epsilon^2})$ (see Lemma 4.7 for lower bound, upper bound comes from Chernoff bound).

- **Condition (4.1) is for all** D. It is not interesting to compare the SL sample complexity and SSL sample complexity over *a* worst-case D. The quantifier must be universal, not existential, otherwise it is not hard to show the conjecture because we can let D to be concentrated on $\mathrm{VC}(\mathcal{H})$ shattered points and both SL and SSL will have same sample complexity (see Corollary 4.10).

Thus, what's interesting about the for all quantifier over D is that it includes D's for which there exists a low density separator, which many practitioners doing SSL believe is a good predictor (it turns out the low density separator can be consistently estimated with unlabeled data (see Ben-David et al., 2009)), but the conjecture says there is some bad labeled distribution $P \in \mathsf{Ext}(D)$ for which SSL provides no significant advantage.

- **Taking the supremum is considering the worst possible labeled distribution.** And perhaps this is the *most unrealistic and controversial* aspect of our analysis. In practice, people make assumptions when doing SSL. These assumptions are about the relationship between the labels (i.e. conditional distribution given $x \in \mathcal{X}$) and the unlabeled distribution. For example, for the cluster assumption, one should expect that the low density region of D divides the domain into different label classes (of course there may still be labelling noise). Thus, rather than taking worst case over $\mathsf{Ext}(D)$ it is perhaps more realistic to take the worst case over extensions of D that are reasonable with respect to the cluster assumption.

There is also a form of this conjecture for the realizable setting. Much of the earlier discussion also applies to this setting.

Conjecture 4.2. *In the realizable setting, there exists constants $C \geq 1$, $\xi, \xi' > 0$, and a fixed empirical risk minimization algorithm ERM_0 such that for all $\mathcal{X}$, $\epsilon \in (0, \xi)$, $\delta \in (0, \xi')$, for every hypothesis class $\mathcal{H}$ over $\mathcal{X}$, for every distribution D over $\mathcal{X}$ we have*

$$\sup_{h \in \mathcal{H}} m_{\mathrm{SL}}(\mathsf{ERM}_0, \epsilon, \delta, \mathcal{H}, D_h) \leq C \min_{\mathcal{A}} \sup_{h \in \mathcal{H}} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, D_h).$$

The results of this chapter are mainly concerned with lower bounds of sample complexity on SSL and upper bounds of sample complexity on SL (usually given by a ERM type algorithm), and whether these match up to a constant factor. Any lower bound on SSL is also a lower bound on SL since, in particular, a SL algorithm is a SSL algorithm that does not use knowledge of D, the unlabeled distribution. What is meant by a lower bound on SSL is that there exists no SSL algorithm (one that must PAC learn according our earlier definitions) that can have sample complexity smaller than the lower bound, not even algorithms that make SSL type assumptions since we are in the no prior knowledge setting.

The upper bounds on SL are given by ERM algorithms, typically as a result of uniform convergence, however, this does not rule out any algorithms that might perform better than ERM in some cases. For example, one might do regularization and if one is lucky then one can do better in some cases. But regularization may not always work well. For example, for some tasks, the optimal linear halfspace has small margin and the regularization algorithm would instead pick a separator with large margin resulting in larger error.

Thus, the real gap in sample complexity between SL and SSL might be smaller than what we prove, but we prove that the gap must be smaller than a multiplicative factor. Note also that when the VC dimension of a hypothesis class is infinity, then it does not make sense to compare SSL and SL in our model as learning cannot be done. Of course, in practical situations people use regularization to compensate for infinite VC classes, but the success of such an approach depends on whether the regularizer is indeed good complexity control. For example, when using margin regularization it is possible that the task at hand does not have good margin to begin with.

In the remainder of this thesis, we will prove the conjecture for some specific cases. Namely for the hypothesis classes of thresholds and union of intervals over the real line. We do not prove the conjecture in its entirety, we obtain partial results where the assertion of

> "for every distribution D over $\mathcal{X}$"

in Conjectures 4.1 and 4.2 is replaced with

> "for every distribution D over $\mathcal{X}$ that has a density function."

Note that in our examples $\mathcal{X} = \mathbb{R}$. First, in the next section we show that SSL over certain natural hypothesis classes over the real line, assuming that the unlabeled distribution is "smooth," is equivalent to supervised learning with respect to the uniform distribution.

4.2 Reduction to the Uniform Distribution on $[0, 1]$

In this section we show that, over some natural hypothesis classes over the real line, semi-supervised learning with respect to any nicely behaving distribution is, perhaps surprisingly, equivalent to supervised (or semi-supervised learning) with respect to the uniform distribution over the unit interval. This reduction simplifies the proofs of upper and lower bounds on the sample complexity of semi-supervised learning as we will only need to consider the uniform distribution. Unfortunately this reduction technique is not useful for certain kinds of unlabeled distributions, for example, ones which put positive probability mass on a few points.

In Section 4.2.1 we informally describe this reduction technique and provide the main idea why it is sufficient to focus on semi-supervised learning on the uniform distribution. We give a formal statement of this technique, its consequences and some proofs in Section 4.2.2.

4.2.1 The Rough Idea

Given any "nice" unlabeled distribution D over $\mathbb{R}$ (to be defined in Section 4.2.2), we can map each point in $\mathbb{R}$ to D's cumulative distribution function $x \mapsto F(x)$. The induced distribution $F(D)$ is over the unit interval and is uniform because for any $t \in \mathbb{R}$ such that $D((-\infty, t]) = p$, we have $F(D)([0,p]) = p$. Figure 4.1 shows this reduction.

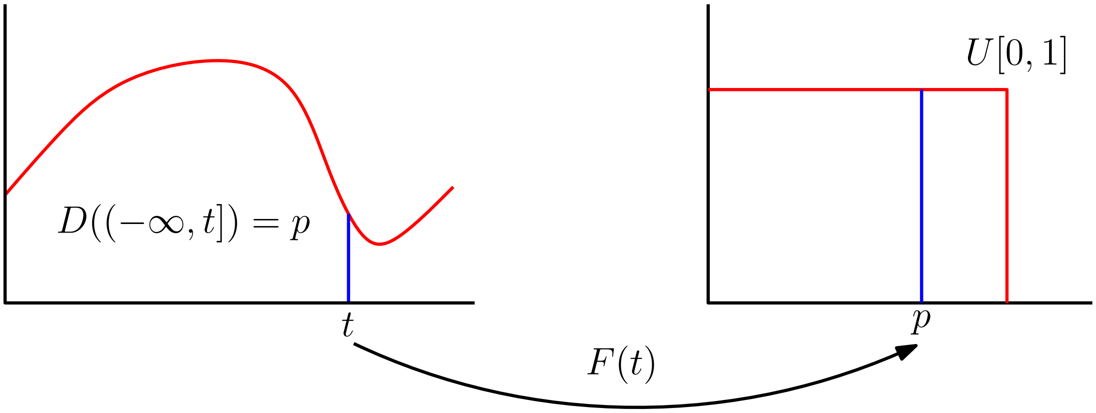

Figure 4.1: One can simplify the problem of semi-supervised learning on any "nice" unlabeled distribution to learning with respect to the uniform distribution on $[0,1]$. The figure shows that the mapping $F(t) = D((-\infty, t])$ induces the uniform distribution.

Given any semi-supervised algorithm $\mathcal{A}$ that "learns" with respect to all $P \in \mathsf{Ext}(U)$ where U is the uniform distribution over $[0,1]$ (i.e. $\mathcal{A}$ is "tailored" for the uniform distribution), it can be turned into a semi-supervised learning algorithm that learns any D (with a few restrictions on its niceness) as follows:

Semi-supervised algorithm $\mathcal{B}(\{(x_i, y_i)\}_{i=1}^m, D)$

1. Compute the cumulative distribution function of D, call it F.
2. Compute $S' = \{(F(x_i), y_i)\}_{i=1}^m$
3. Output $\mathcal{A}(S') \circ F$.

The last line essentially converts hypotheses that are restricted to the unit interval to hypotheses over $\mathbb{R}$.

In the other direction, we need to show that it is not advantageous to have access to any unlabeled distribution ("nice" one) compared to learning on the uniform distribution. Fix D, if we are given any semi-supervised algorithm $\mathcal{B}$, then we can create a semi-supervised algorithm $\mathcal{A}$ for the uniform distribution that has at most the sample complexity of $\mathcal{B}$.

Semi-supervised algorithm $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^m, U)$

1. Let F^{-1} be the inverse cumulative distribution function of D
2. Compute $S' = \{(F^{-1}(x_i), y_i)\}_{i=1}^m$
3. Output $\mathcal{B}(S', U) \circ F^{-1}$.

The above algorithm $\mathcal{A}$ is designed for any $\mathcal{B}$ and any "nice" D. Thus, the sample complexity of any SSL algorithm for any "nice" D is equivalent to the sample complexity of any SSL algorithm for the uniform distribution. See Section 4.2.2 for a formal statement of this fact.

Thus, to prove that semi-supervised learning does not give much advantage compared to supervised learning, we are essentially showing that

> *Distribution-free learning (supervised learning setting) is more or less as hard as learning with respect to the uniform distribution.*

By proving a sample complexity lower bound on learning under the uniform distribution, we are really proving a lower bound on the sample complexity of semi-supervised learning under any nice unlabeled distribution. In Section 4.3 and 4.4 we make use of this reduction to the uniform distribution in conjunction with an averaging argument to obtain lower bounds on the semi-supervised learning sample complexity that is within a constant factor of well-known upper bounds on supervised learning sample complexity (e.g. distribution-free VC-dimension upper bounds on the sample complexity of ERM).

4.2.2 Formal Proof

Let us start by defining the hypothesis classes. Recall the class of thresholds is defined as $H = \{\mathbf{1}(-\infty, t] \, : \, t \in \mathbb{R}\}$ and the class of union of d intervals is

$$\mathrm{UI}_d := \{\mathbf{1}[a_1, a_2) \cup [a_3, a_4) \cup \cdots \cup [a_{2\ell-1}, a_{2\ell}) \; : \; \ell \leq d, \; a_1 \leq a_2 \leq \cdots \leq a_{2\ell}\} \, .$$

The reduction to the uniform distribution simplification says that the semi-supervised learning sample complexity of learning H (or UI_d) under any unlabeled distribution (having a probability density) is equivalent to the sample complexity of a supervised learning algorithm "tailored" for the unlabeled uniform distribution (say on the unit interval). In other words, assuming the unlabeled distribution has density, they are all essentially the same for semi-supervised learning algorithms.

In this section we formally prove that learning any "natural" hypothesis class on the real line has the same sample complexity for any unlabeled distribution (having density) and is independent of its "shape." Intuitively, if we imagine the real axis made of rubber, then a natural hypothesis class is one that is closed under stretching of the axis. Classes of thresholds and union of d intervals are examples of such natural classes, since under any rescaling an interval remains an interval.

The rescaling will apply also on the unlabeled distribution over the real line and it will allow us to go from any unlabeled distribution (having density) to the uniform distribution over $(0, 1)$.

Definition 4.1. A *rescaling* function $f : \mathbb{R} \to [0, 1]$ is a function that is continuous and increasing.

Definition 4.2. Fix a hypothesis class $\mathcal{H}$ and let $\mathcal{H}|_{[0,1]} = \{h|_{[0,1]} : h \in \mathcal{H}\}$ where $h_{[0,1]}$ is the restriction of h to the domain $[0, 1]$. A hypothesis class $\mathcal{H}$ is *closed under rescaling* f if for all $h \in \mathcal{H}$, we have that $\mathbf{1}f(\mathsf{set}(h)) = h \circ f \in \mathcal{H}|_{[0,1]}$.

Our technical results in the next few sections are for the continuous uniform distribution on $[0, 1]$. Thus, we shall limit ourselves to unlabeled distributions that under rescaling, is equal to the uniform distribution. Such classes of distributions have a probability density (as defined below). For example, distributions that have positive probability on a single point does not belong to the class.

Definition 4.3. A distribution D over $\mathbb{R}$ has *probability density* g with respect to the Lebesgue measure if for any Lebesgue measurable set A,

$$D(A) = \int_A g(x)\ \mathrm{d}x$$

where the integral is taken with respect to the Lebesgue measure.

The Radon–Nikodym theorem says that, on the real line, distributions with probability densities are precisely those distributions that are *absolutely continuous*. In this thesis we will refer to these distributions as *having probability densities* or simply having density in order to give a more intuitive understanding for those not familiar with measure theory. Note that this is in contrast with our terminology in Ben-David et al. (2008). The following definition is reproduced for completeness. Our rescalings will be defined by the cumulative distribution, which when it has density, is continuous.

Definition 4.4. For a distribution D over $\mathbb{R}$, the *cumulative distribution function* F is

$$F(x) = D((-\infty, x]).$$

If D has density f, then $F(x) = \int_{-\infty}^{x} f(x)\ \mathrm{d}x$.

Example 4.1. Let D be any unlabeled distribution having density, then the cumulative distribution function F of D is continuous and strictly increasing. Thus F is a rescaling. The class of thresholds and the class of unions of d intervals are closed under rescaling F. This can be observed from the fact that for any interval I, $F(I)$ is an interval in $[0, 1]$.

We show that the sample complexity is unaffected by the rescalings provided that the hypothesis class is closed under rescalings. We split the results into two lemmas—Lemma 4.1 and Lemma 4.2. The first lemma shows that if we have a supervised algorithm with certain sample complexity for the case when the unlabeled distribution is the uniform distribution over $[0,1]$, then the algorithm can be translated into a semi-supervised algorithm with the same sample complexity for the case when the unlabeled distribution has density. The second lemma shows the translation in the other direction. Namely, that a semi-supervised algorithm with certain sample complexity on some unlabeled distribution (having density) can be translated to a supervised algorithm for the case when unlabeled distribution is uniform over $[0,1]$.

Lemma 4.1. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}$ closed under rescaling. Let U be the uniform distribution over $(0,1)$. Let $\epsilon, \delta > 0$.*

(a) (Realizable case)*: If $\mathcal{A}$ is any semi-supervised algorithm, then there exists a semi-supervised learning algorithm $\mathcal{B}$ such that for any distribution D over $\mathbb{R}$ which has density (with respect to the Lebesgue measure)*

$$\sup_{h \in \mathcal{H}} m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, D_h) \leq \sup_{g \in \mathcal{H}} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, U_g) \ . \tag{4.2}$$

(b) (Agnostic case)*: If $\mathcal{A}$ is any semi-supervised algorithm, then there exists a semi-supervised learning algorithm $\mathcal{B}$ such that for any distribution D over $\mathbb{R}$ which has density (with respect to the Lebesgue measure)*

$$\sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, P) \leq \sup_{Q \in \mathsf{Ext}(U)} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, Q) \ . \tag{4.3}$$

Proof. We prove this lemma for the agnostic case, but can be easily extended to the realizable case (by replacing some quantities).

Fix $\mathcal{H}$ and $\mathcal{A}$. We construct algorithm $\mathcal{B}$ as follows. The algorithm $\mathcal{B}$ has two inputs, a sample $S = \{(x_i, y_i)\}_{i=1}^m$ and a distribution D. Based on D the algorithm computes the cumulative distribution function $F(t) = D((-\infty, t])$. Then, $\mathcal{B}$ computes from S the transformed sample $S' = \{(x'_i, y_i)\}_{i=1}^m$ where $x'_i = F(x_i)$. On a sample S' and distribution U, the algorithm $\mathcal{B}$ simulates algorithm $\mathcal{A}$ and computes $h = \mathcal{A}(S')$. Finally, $\mathcal{B}$ outputs $g = h \circ F$.

It remains to show that for any D with continuous cumulative distribution function (4.2) and (4.3) holds for any $\epsilon, \delta > 0$. We prove (4.3), the other equality is proved similarly.

Let $P \in \mathsf{Ext}(D)$. Slightly abusing notation, we define the "image" distribution $F(P)$ over $(0,1) \times \{0,1\}$ to be

$$F(P)(M) = P(\{(x,y) \ : \ (F(x), y) \in M\})$$

for any (Lebesgue measurable) $M \subseteq (0,1) \times \{0,1\}$. It is not hard to see that if S is distributed according to P^m, then S' is distributed according to $(F(P))^m$. Since D has density and therefore F is continuous, it can be seen that $\mathcal{D}(F(P)) = U$

(i.e. $F(P) \in \mathsf{Ext}(U)$). Further F is a rescaling (since it is also increasing). Hence $\mathrm{Er}^{F(P)}(h) = \mathrm{Er}^{P}(h \circ F)$ and $\inf_{h \in H} \mathrm{Er}^{P}(h) = \inf_{h \in H} \mathrm{Er}^{F(P)}(h)$. We have for any ϵ and any $m \geq 0$

$$\begin{aligned}
&\Pr_{S \sim P^m}[\mathrm{Er}^{P}(\mathcal{B}(S, D)) - \inf_{h \in H} \mathrm{Er}^{P}(h) > \epsilon] \\
&= \Pr_{S' \sim F(P)^m}[\mathrm{Er}^{P}(\mathcal{A}(S') \circ F) - \inf_{h \in H} \mathrm{Er}^{F(P)}(h) > \epsilon] \\
&= \Pr_{S' \sim F(P)^m}[\mathrm{Er}^{F(P)}(\mathcal{A}(S')) - \inf_{h \in H} \mathrm{Er}^{F(P)}(h) > \epsilon] \ .
\end{aligned}$$

Therefore, for any $\epsilon, \delta > 0$,

$$\begin{aligned}
m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, P) &= m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, F(P)) \\
&\leq \sup_{Q \in \mathsf{Ext}(U)} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, Q) \ .
\end{aligned} \tag{4.4}$$

Taking supremum over $P \in \mathsf{Ext}(D)$ over Equation (4.4) finishes the proof. □

Lemma 4.2. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}$ closed under rescaling. Let U be the uniform distribution over $[0, 1]$. Let $\epsilon, \delta > 0$.*

(a) (Realizable case)*: If $\mathcal{B}$ is any semi-supervised algorithm and D is any distribution over $\mathbb{R}$ which has density (with respect to the Lebesgue measure), then there exists a semi-supervised algorithm $\mathcal{A}$ such that*

$$\sup_{g \in \mathcal{H}} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, U_g) \leq \sup_{h \in \mathcal{H}} m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, D_h) \ . \tag{4.5}$$

(b) (Agnostic case)*: If $\mathcal{B}$ is any semi-supervised algorithm and D is any distribution over $\mathbb{R}$ which has density (with respect to the Lebesgue measure), then there exists a semi-supervised algorithm $\mathcal{A}$ such that*

$$\sup_{Q \in \mathsf{Ext}(U)} m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, Q) \leq \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, P) \ . \tag{4.6}$$

Proof. We prove the lemma for the agnostic case since the realizable case is similar.

Fix $\mathcal{H}$, $\mathcal{B}$ and D. Let F be the be cumulative distribution function of D. Since D has density, F is a rescaling and inverse F^{-1} exists.

Now, we construct algorithm $\mathcal{A}$. Algorithm $\mathcal{A}$ maps input sample $S' = \{(x_i', y_i)\}_{i=1}^m$ to sample $S = \{(x_i, y_i)\}_{i=1}^m$ where $x_i = F^{-1}(x_i')$. On a sample S the algorithm $\mathcal{A}$ simulates algorithm $\mathcal{B}$ and computes $g = \mathcal{B}(S, D)$. Finally, $\mathcal{A}$ outputs $h = g \circ F^{-1}$.

It remains to show that for any D with continuous cumulative distribution function (4.5) and (4.6) holds for any $\epsilon, \delta > 0$. We prove (4.6), the other equality is proved similarly.

Let $Q \in \mathsf{Ext}(U)$. Slightly abusing notation, we define the "pre-image" distribution $F^{-1}(Q)$ over $I \times \{0, 1\}$ to be

$$F^{-1}(Q)(M) = Q\left(\{(F(x), y) \ : \ (x, y) \in M\}\right)$$

for any (Lebesgue measurable) $M \subseteq I \times \{0,1\}$. It is not hard to see that if S' is distributed according to Q, then S is distributed according to $(F^{-1}(Q))^m$. We have $\mathcal{D}(F^{-1}(U)) = D$ (i.e. $F^{-1}(Q) \in \mathsf{Ext}(D)$). Since F^{-1} is a rescaling, $\mathrm{Er}^{F^{-1}(Q)}(h) = \mathrm{Er}^{Q}(h \circ F^{-1})$ and $\inf_{h \in H} \mathrm{Er}^{Q}(h) = \inf_{h \in H} \mathrm{Er}^{F^{-1}(Q)}(h)$. We have for any $\epsilon > 0$ and any $m \in \mathbb{N}$

$$\begin{aligned}
&\Pr_{S' \sim Q^m}[\mathrm{Er}^{Q}(\mathcal{A}(S')) - \inf_{h \in H} \mathrm{Er}^{Q}(h)] \\
&= \Pr_{S \sim F^{-1}(Q)^m}[\mathrm{Er}^{Q}(\mathcal{B}(S,D) \circ F^{-1}) - \inf_{h \in H} \mathrm{Er}^{F^{-1}(Q)}(h)] \\
&= \Pr_{S \sim F^{-1}(Q)^m}[\mathrm{Er}^{F^{-1}(Q)}(\mathcal{B}(S,D)) - \inf_{h \in H} \mathrm{Er}^{F^{-1}(Q)}(h)] \,.
\end{aligned}$$

Therefore, for any $\epsilon, \delta > 0$,

$$\begin{aligned}
m_{\mathrm{SL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, Q) &= m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, F^{-1}(Q)) \\
&\le \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{B}, \epsilon, \delta, \mathcal{H}, P).
\end{aligned} \tag{4.7}$$

Taking supremum over $Q \in \mathsf{Ext}(U)$ in Equation (4.7) finishes the proof. □

4.3 Learning Thresholds in the Realizable Setting

In this section we consider learning the class of thresholds, $H = \{\mathbf{1}(-\infty, t] : t \in \mathbb{R}\}$, on the real line in the realizable setting and show that for any unlabeled distribution (having density) semi-supervised learning has at most a factor of 2 advantage over supervised learning in the sample complexity, taken over worst case distributions on $\mathcal{X} \times \mathcal{Y}$ whose marginal over $\mathcal{X}$ is equal to D.

First, in Theorem 4.3, we reproduce a $\frac{\ln(1/\delta)}{\epsilon}$ upper bound on the sample complexity of supervised learning. This is a well-known "textbook" result in computational learning theory. Second, we consider the sample complexity of semi-supervised learning in the case when the unlabeled distribution has density with respect to the Lebesgue measure on $\mathbb{R}$ (since we use the reduction technique in Section 4.2). In Theorems 4.4 and 4.5 we show that the sample complexity is between $\frac{\ln(1/\delta)}{2\epsilon} + O(\frac{1}{\epsilon})$ and $\frac{\ln(1/\delta)}{2.01\,\epsilon} - O(\frac{1}{\epsilon})$.[2] Ignoring the lower order terms, we see that the sample complexity of supervised learning is (asymptotically) at most 2-times larger than that of semi-supervised learning.

In the remainder of this section, we will be analyzing the sample complexity of the supervised learning algorithm LeftERM that chooses the right most positive point as the threshold (for precise definition, see Equation 2.3) and the semi-supervised algorithm MedianERM proposed by Kääriäinen (2005) which is defined as follows.

[2] The 2.01 in the lower bound can be replaced by arbitrary number strictly greater than 2. This slight imperfection is a consequence of that the true dependence of the sample complexity on ϵ, in this case, is of the form $1/\ln(1-2\epsilon)$ and not $1/(2\epsilon)$.

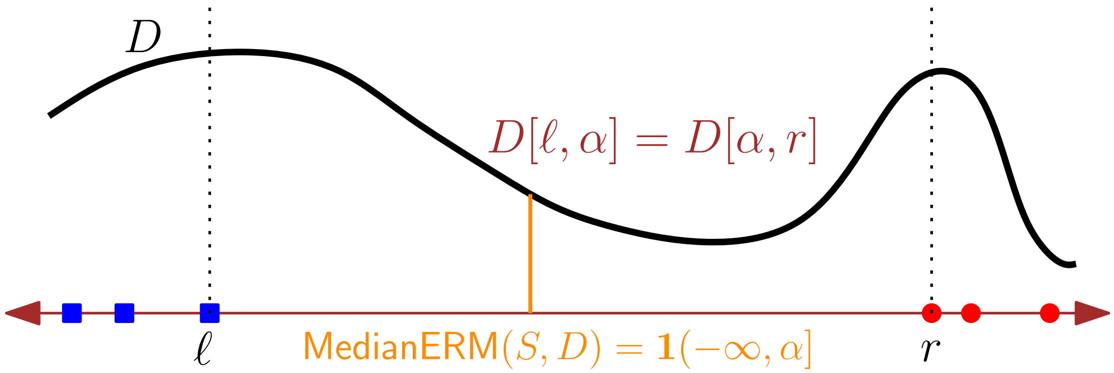
D
$D[\ell, \alpha] = D[\alpha, r]$
ℓ
$\mathsf{MedianERM}(S, D) = \mathbf{1}(-\infty, \alpha]$
r

Figure 4.2: Given a labeled sample S (squares are '1', circles are '0') and an unlabeled distribution D, $\mathsf{MedianERM}(S, D)$ first computes ℓ and r, the right most '1' and the left most '0', respectively. Then it outputs the threshold at α such that $D[\ell, \alpha] = D[\alpha, r]$.

Definition 4.5 (Algorithm $\mathsf{MedianERM}$)**.** Let D be any unlabeled distribution and

$$S = \{(x_1, y_2), (x_2, y_2), \ldots, (x_m, y_m)\}$$

be labeled training data. Let

$$\begin{aligned} \ell &= \max\{x_i \ : \ i \in [m], \ y_i = 1\} \ , \\ r &= \min\{x_i \ : \ i \in [m], \ y_i = 0\} \ , \\ t &= \sup\{a \ : \ D((\ell, t']) \leq D((\ell, r])/2\} \ , \end{aligned}$$

then (see Definition 3.2 for semi-supervised algorithm)

$$\mathsf{MedianERM}(S, D) = \mathbf{1}(-\infty, t] \ .$$

That is, the algorithm outputs the *median* of the distribution D restricted on the interval $[\ell, r)$. See Figure 4.2 for an illustration of this algorithm.

In general, $\mathsf{MedianERM}$ can be extended to arbitrary hypothesis classes for the realizable setting (see Kääriäinen, 2005). The idea is to define a pseudo-metric space[3] $(\mathcal{H}, d_D)$ where D is the input unlabeled distribution such that for $h_1, h_2 \in \mathcal{H}$,

$$d_D(h_1, h_2) = D\{\mathsf{set}(h_1) \Delta \, \mathsf{set}(h_2)\}.$$

Then on a sample S the algorithm considers the consistent hypotheses $V = \{h \in \mathcal{H} \ : \ \mathrm{Er}^S(h) = 0\}$, and chooses the hypothesis

$$\hat{h} = \operatorname*{argmin}_{f \in V} \max_{g \in V} d_D(f, g).$$

[3] A pseudo-metric space is a metric space in which the distance between two distinct points in the domain can be zero.

Clearly this is an ERM paradigm, and intuitively one expects this to do twice as good as a supervised learning algorithm that does not have access to D and therefore may, in the worst case, choose a hypothesis that lies on the "boundary" of V whereas the semi-supervised algorithm would choose the "centre." But it is unknown whether this will always provide a factor two advantage in the sample complexity over supervised ERM algorithms as it depends on the structure of $\mathcal{H}$.

In our analysis below, we show that it appears to give a factor two advantage in sample complexity over supervised learning algorithms when $\mathcal{H}$ are thresholds. It remains an open question what the lower bound for supervised learning sample complexity is for thresholds in the realizable setting, although we strongly believe it is factor two worse than semi-supervised learning (e.g. one can show $\mathsf{LeftERM}$ in the worse case is twice worse than $\mathsf{MedianERM}$ on learning the target $\mathbf{1}(0,1]$ when $\mathcal{X} = (0,1]$).

Theorem 4.3 (Supervised Learning Upper Bound)**.** *Let H be the class of thresholds and* $\mathsf{LeftERM}$ *be the supervised learning algorithm as in Definition 2.3. For any D, for any $\epsilon, \delta > 0$,*

$$\sup_{h \in H} m_{\mathrm{SL}}(\mathsf{LeftERM}, \epsilon, \delta, H, D_h) \leq \frac{\ln(1/\delta)}{\epsilon} .$$

Proof. This is a well-known result and can be found in textbooks such as Kearns and Vazirani (1994), we prove it here for convenience. Suppose we fix a "target" h, and $\epsilon, \delta > 0$. For a fixed labeled sample S that is consistent with h, the event that $\mathrm{Er}^P(\mathsf{LeftERM}(S)) \leq \epsilon$ occurs if and only if the event (with a little abuse of notation) $S \cap [t - \xi, t] \neq \emptyset$ occurs where $\xi = \inf\{a \leq t : \mathcal{D}(P)([a,t]) \leq \epsilon\}$ (that is, some point hits the interval $[t - \xi, t]$). The probability that a point hits the interval $[t - \xi, t]$ is exactly

$$\Pr_{S \sim P^m}(\mathrm{Er}^P(\mathsf{LeftERM}(S)) \leq \epsilon) = 1 - (1 - \epsilon)^m$$

Thus, we require that

$$\begin{aligned} \delta &\leq (1 - \epsilon)^m \\ m &\geq \frac{\ln \delta}{\ln(1 - \epsilon)} \\ &\geq \frac{\ln \delta}{-\epsilon} && \text{since } \exp(x) \geq 1 + x \text{ for } x \leq 1. \\ &= \frac{\ln \frac{1}{\delta}}{\epsilon}. && \square \end{aligned}$$

Now we show that $\mathsf{MedianERM}$ can take advantage of the knowledge of the distribution by picking the unlabeled distribution's median in the version space. The sample complexity indeed drops by a factor of two compared with $\mathsf{LeftERM}$.

Theorem 4.4 (Semi-Supervised Learning Upper Bound)**.** *Let H be the class of thresholds and* $\mathsf{MedianERM}$ *be the semi-supervised learning algorithm as in Definition 4.5. For any unlabeled distribution D having density, any $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{2})$, and any "target" $h \in H$,*

$$m_{\text{SSL}}(\mathsf{MedianERM}, \epsilon, \delta, H, D_h) \leq \frac{\ln(1/\delta)}{2\epsilon} + \frac{\ln 2}{2\epsilon} \; .$$

Proof. By reduction to the uniform distribution (see Lemma 4.1 part (a)) we can assume that D is uniform over $[0, 1]$. Fix $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{2})$ and $h \in H$. We show that, for any $m \geq 2$,

$$\Pr_{S \sim D_h^m}[\mathrm{Er}^{D_h}(\mathsf{MedianERM}(S, D)) \geq \epsilon] \leq 2(1 - 2\epsilon)^m \; , \tag{4.8}$$

from which the theorem easily follows, since if $m \geq \frac{\ln(1/\delta)}{2\epsilon} + \frac{\ln 2}{2\epsilon}$, then $m \geq 2$ and $2(1 - 2\epsilon)^m \leq 2 \exp(-2m\epsilon) \leq \delta$.

In order to prove (4.8), let $h = \mathbf{1}(-\infty, t]$ be the "target". Without loss of generality $t \in [0, \frac{1}{2}]$. Let us assume that $\ell \in [0, t]$ and $r \in [t, 1]$. When given S, ℓ, r and D the uniform distribution, then $\mathsf{MedianERM}(S, D) = (\ell + r)/2$ the midpoint. For $\mathrm{Er}^{D_h}(\mathsf{MedianERM}(S, D)) \leq \epsilon$ to hold, we require

$$\begin{aligned} t - \epsilon \leq \frac{\ell + r}{2} &\leq t + \epsilon \\ \iff r \in [\max(2t - \ell - 2\epsilon, t), &\min(2t - \ell + 2\epsilon, 1)] \end{aligned}$$

Let $a(\ell) = \max(2t - \ell - 2\epsilon, t)$ and $b(\ell) = \min(2t - \ell + 2\epsilon, 1)$. We lower bound the probability of success

$$p = \Pr_{S \sim D_h^m}[\mathrm{Er}^{D_h}(\mathsf{MedianERM}(S, D)) \leq \epsilon] \; .$$

There are two cases:

Case 1: If $t > 2\epsilon$, then we integrate over all possible choices of the rightmost positive example in S (which determines ℓ) and leftmost negative example in S (which determines r). There are $m(m-1)$ choices for the rightmost positive example and leftmost negative example. Also, we require that no other points (i.e. $m - 2$ of them) fall in between $[\ell, r]$, which occurs with probability $(1 - (r - \ell))^{m+2}$. We have

$$p \geq p_1 = m(m-1) \int_0^t \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r \mathrm{d}\ell \; .$$

This inequality is actually strict since we have not accounted for the case when no positive points are in the sample, but the output can still be ϵ close to t.

Case 2: If $t \leq 2\epsilon$, then we integrate over all possible choices of the rightmost positive example in S and leftmost negative example in S. Additionally we also consider samples without positive examples (the second term in the sum), and integrate over all possible choices of the leftmost (negative) example. We have

$$p \geq p_2 = m(m-1) \int_0^t \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell + m \int_t^{2\epsilon} (1-r)^{m-1} \, \mathrm{d}r$$

Both cases split into further subcases.

Subcase 1a: If $t > 2\epsilon$ and $t + 4\epsilon \leq 1$ and $t + \epsilon \geq 1/2$, then $0 \leq 2t + 2\epsilon - 1 \leq t - 2\epsilon \leq t$ and

$$\begin{aligned}
p_1 &= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t-2\epsilon} \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{2t-\ell-2\epsilon}^{1} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t-2\epsilon} \int_{2t-\ell-2\epsilon}^{2t-\ell+2\epsilon} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{t}^{2t-\ell+2\epsilon} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= 1 - \frac{1}{2}(1-2t-2\epsilon)^m - \frac{1}{2}(-1+2t+6\epsilon)^m - (1-2\epsilon)^m \\
&\geq 1 - 2(1-2\epsilon)^m \, .
\end{aligned}$$

Subcase 1b: If $t > 2\epsilon$ and $t + \epsilon \leq 1/2$, then $2t + 2\epsilon - 1 \leq 0 \leq t - 2\epsilon \leq t$ and

$$\begin{aligned}
p_1 &= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{a(\ell)}^{b(\ell)} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{2t-\ell-2\epsilon}^{2t-\ell+2\epsilon} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{t}^{2t-\ell+2\epsilon} (1-r+\ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= 1 - (1-2\epsilon)^m + \frac{1}{2}(1-2t-2\epsilon)^m - \frac{1}{2}(1-2t+2\epsilon)^m \\
&\geq 1 - \frac{3}{2}(1-2\epsilon)^m \, .
\end{aligned}$$

Subcase 1c: If $t > 2\epsilon$ and $t + 4\epsilon \geq 1$, then $0 \leq t - 2\epsilon \leq 2t + 2\epsilon - 1 \leq t$, and

$$\begin{aligned}
p_1 &= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{2t+2\epsilon-1} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{2t-\ell-2\epsilon}^{1} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{2t+2\epsilon-1} \int_{t}^{1} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t} \int_{t}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= 1 - (1-2\epsilon)^m - \frac{1}{2}(1 - 2t + 2\epsilon)^m - \frac{1}{2}(2t + 2\epsilon - 1)^m \\
&\geq 1 - 2(1-2\epsilon)^m \,.
\end{aligned}$$

Subcase 2a: If $t \leq 2\epsilon$ and $t + \epsilon \geq 1/2$, then $t - 2\epsilon \leq 0 \leq 2t + 2\epsilon - 1 \leq t$ and

$$\begin{aligned}
p_2 &= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&\quad + m \int_t^{2\epsilon} (1-r)^{m-1} \, \mathrm{d}r \\
&= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{t}^{1} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t} \int_{t}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&\quad + (1-t)^m - (1-2\epsilon)^m \\
&= 1 - \frac{3}{2}(1-2\epsilon)^m - \frac{1}{2}(2t + 2\epsilon - 1)^m \\
&\geq 1 - 2(1-2\epsilon)^m \,.
\end{aligned}$$

Subcase 2b: If $t \leq 2\epsilon$ and $t + \epsilon \leq 1/2$, then $t - 2\epsilon \leq 0$, $2t + 2\epsilon - 1 \leq 0$ and

$$\begin{aligned}
p_2 = m(m-1) \int_0^t \int_{a(\ell)}^{b(\ell)} & (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
& + m \int_t^{2\epsilon} (1-r)^{m-1} \, \mathrm{d}r \\
& = m(m-1) \int_0^t \int_t^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
& + (1-t)^m - (1-2\epsilon)^m \\
& = 1 - \frac{3}{2}(1-2\epsilon)^m - \frac{1}{2}(1 - 2t - 2\epsilon)^m \\
& \geq 1 - 2(1-2\epsilon)^m \; .
\end{aligned}$$

We could have removed subcase 1c by assuming $\epsilon \in (0, 1/8)$, for example. □

Finally, we can also show that MedianERM is the optimal semi-supervised learning algorithm up to a constant additive factor of $O(1/\epsilon)$. Note that this lower bound also applies to supervised learning, since any supervised learning algorithm can be turned into a semi-supervised learning algorithm by just ignoring the input unlabeled distribution.

Theorem 4.5 (Semi-Supervised Learning Lower Bound)**.** *For any randomized semi-supervised algorithm $\mathcal{A}$ that can output stochastic hypotheses, any $\epsilon \in (0, 0.001)$, any $\delta > 0$, any unlabeled distribution D (with density), there exists $h \in H$, such that*

$$m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, H, D_h) \geq \frac{\ln(1/\delta)}{2.01\epsilon} - \frac{\ln 2}{2.01\epsilon} \; .$$

Proof. By reduction to the uniform distribution (see Lemma 4.2 part (a)) we can assume that D is uniform over $[0, 1]$. Fix $\mathcal{A}, \epsilon, \delta$. For a fixed labeled sample S, let Q_S be the distribution over all functions $\{0,1\}^[0, 1]$ such that $\mathcal{A}(S, D) \sim Q_S$ (for a deterministic $\mathcal{A}$, the support of Q_S is a single hypothesis). We show the existence of a "bad" h by an averaging argument. Let t be a random variable uniformly distributed on $[0, 1]$ and let $h = \mathbf{1}(-\infty, t]$. We prove that for all $m \geq 0$,

$$\mathop{\mathbb{E}}_{t} \Pr_{S \sim D_h^m} \mathop{\mathbb{E}}_{\mathcal{A}(S,D) \sim Q_S} [\mathrm{Er}^{D_h}(\mathcal{A}(S, D)) \geq \epsilon] \geq \frac{1}{2}(1 - 2\epsilon)^m \; . \tag{4.9}$$

For the remainder of this proof, we will simplify notation by ignoring the expectation over Q_S (we show later it does not matter). The left-hand side can rewritten as

$$\mathop{\mathbb{E}}_{t} \Pr_{S \sim D_h^m} [\mathrm{Er}^{D_h}(\mathcal{A}(S, D)) \geq \epsilon] = \mathop{\mathbb{E}}_{t} \mathop{\mathbb{E}}_{S \sim D_h^m} \mathbf{1}\{(t, S) \; : \; \mathrm{Er}^{D_h}(\mathcal{A}(S, D)) \geq \epsilon\}.$$

For t chosen from the uniform distribution on $(0,1)$ and then drawing $S \sim D_h^m$, S has the same distribution as drawing $S' \sim D^m$ and labelling S' with a random t. Hence, we can flip the expectations on the right hand side to get

$$\begin{aligned} &= \mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{t} \mathbf{1}\{(t,S) \ : \ \mathrm{Er}^{D_h}(\mathcal{A}((S,h(S)),D)) \geq \epsilon\} \\ &= \mathop{\mathbb{E}}_{S \sim D^m} \Pr_{t}[\mathrm{Er}^{D_h}(\mathcal{A}((S,h(S)),D)) \geq \epsilon] \end{aligned}$$

To lower bound the last expression, fix unlabeled points $0 \leq x_1 \leq x_2 \leq \cdots \leq x_m \leq 1$. For convenience, let $x_0 = 0$ and $x_{m+1} = 1$. We claim that

$$\Pr_t \left[\mathrm{Er}^{D_h}(\mathcal{A}((S,h(S)),D)) \geq \epsilon\right] \geq \sum_{i=0}^{m} \max(x_{i+1} - x_i - 2\epsilon, 0) \ . \tag{4.10}$$

To prove it we fix $i \in [m]$ and restrict t to lie in the interval $(x_i, x_{i+1}]$. The labels in $(S, h(S))$ are hence fixed. Let $g = \mathcal{A}((S,h(S)),D) \sim Q_{(S,h(S))}$, the following holds for any g

$$\int_{x_i}^{x_{i+1}} \mathbf{1}\left\{t \ : \ \mathrm{Er}^{D_h}(g) \geq \epsilon\right\} \mathrm{d}t \geq \max(x_{i+1} - x_i - 2\epsilon, 0) \ , \tag{4.11}$$

which follows from the fact that $\{t \ : \ \mathrm{Er}^{D_h}(g) < \epsilon\}$ is contained in an interval of length at most 2ϵ. Taking the expectation over the right hand side over random choices of g does not affect the quantity since it is independent of g. Summing over all i we obtain (4.10). We note that (4.11) also holds if g is a stochastic function, the details are messier.

In order to prove (4.9) we will compute expectation over $S \sim D^m$ of both sides of (4.10). Expectation of the left side of (4.10) equals to the left side of (4.9). The expectation of the right side of (4.10) is equal to

$$I_m = m! \underbrace{\int_0^{x_{m+1}} \int_0^{x_m} \int_0^{x_{m-1}} \cdots \int_0^{x_2}}_{m \text{ times}} \sum_{i=0}^{m} \max(x_{i+1} - x_i - 2\epsilon, 0) \\ \mathrm{d}x_1 \cdots \mathrm{d}x_{m-2} \mathrm{d}x_{m-1} \mathrm{d}x_m \ ,$$

since there are $m!$ equiprobable choices for the order of the points $x_1, x_2, \ldots, x_m$ among which we choose, without loss of generality, the one with $x_1 \leq x_2 \leq \cdots \leq x_m$. We look at I_m as a function of x_{m+1} and we prove that

$$I_m(x_{m+1}) = (\max(x_{m+1} - 2\epsilon, 0))^{m+1} \ , \tag{4.12}$$

for any $m \geq 0$ and any $x_{m+1} \in [0,1]$. The bound (4.9) follows from (4.12), since $I_m = I_m(1) = (1-2\epsilon)^{m+1} \geq \frac{1}{2}(1-2\epsilon)^m$ for $\epsilon \leq 1/4$. In turn, (4.12) follows, by induction on m, from the recurrence

$$I_m(x_{m+1}) = m \int_0^{x_{m+1}} I_{m-1}(x_m) + x_m^{m-1} \max(x_{m+1} - x_m - 2\epsilon, 0) \ \mathrm{d}x_m \ ,$$

which is valid for all $m \geq 1$. In the base case, $m = 0$, $I_0(x_1) = \max(x_1 - 2\epsilon, 0)$ trivially follows by definition. In the inductive case, $m \geq 1$, we consider two cases. First case, $x_{m+1} < 2\epsilon$, holds since $\max(x_{i+1} - x_i - 2\epsilon, 0) = 0$ and hence by definition $I_m(x_{m+1}) = 0$. In the second case, $x_{m+1} \geq 2\epsilon$, from the recurrence and the induction hypothesis we have

$$\begin{aligned} I_m(x_{m+1}) = m \int_0^{x_{m+1}} & (\max(x_m - 2\epsilon, 0))^m \\ & + \max(x_{m+1} - x_m - 2\epsilon, 0) \cdot x_m^{m-1} \, \mathrm{d}x_m \\ = m & \int_{2\epsilon}^{x_{m+1}} (x_m - 2\epsilon)^m \, \mathrm{d}x_m \\ & + m \int_0^{x_{m+1}-2\epsilon} (x_{m+1} - x_m - 2\epsilon) x_m^{m-1} \, \mathrm{d}x_m \\ = & \frac{m}{m+1}(x_{m+1} - 2\epsilon)^{m+1} \\ & + \frac{1}{m+1}(x_{m+1} - 2\epsilon)^{m+1} \\ = & (x_{m+1} - 2\epsilon)^{m+1} \; . \end{aligned}$$

To finish the proof, suppose $m < \frac{\ln(1/\delta)}{2.01\epsilon} - \frac{\ln 2}{2.01\epsilon}$. Then $\frac{1}{2}(1 - 2\epsilon)^m > \delta$, since

$$\ln\left(\frac{1}{2}(1 - 2\epsilon)^m\right) = -\ln 2 + m \ln(1 - 2\epsilon) > -\ln 2 - m(2.01\epsilon) > \ln \delta \; ,$$

where we have used that $\ln(1 - 2\epsilon) > -2.01\epsilon$ for any $\epsilon \in (0, 0.001)$. Therefore since the average over all targets is at least (4.9), there exists a target $h = \mathbf{1}(-\infty, t]$ such that with probability greater than δ, algorithm $\mathcal{A}$ fails to output a hypothesis with error less than ϵ. □

It is unknown if any supervised algorithm (deterministic or randomized) that has asymptotic sample complexity $c\frac{\ln(1/\delta)}{\epsilon}$ for any constant $c < 1$. For example, perhaps surprisingly, the randomized algorithm RandomERM (see Equation 2.4) that outputs with probability 1/2 the hypothesis $\mathbf{1}(-\infty, \ell]$ and with probability 1/2 the hypothesis $\mathbf{1}(-\infty, r)$ still cannot achieve the semi-supervised learning sample complexity.

We now turn our attention to the agnostic setting, in which we show (limiting to "nice" unlabeled distributions) that semi-supervised learning cannot provide more than a constant factor advantage over supervised learning on the class of thresholds and the union of the intervals.

4.4 Thresholds and Union of Intervals in the Agnostic Setting

In this section, we show that even in the agnostic setting semi-supervised learning does not have more than a constant factor improvement over supervised learning.

We prove some lower bounds for some basic classes over the real line. We introduce the notion of a b-shatterable distribution, which intuitively, are unlabeled distributions where there are b "clusters" that can be shattered by the hypothesis class. The main lower bound of this section are for such distributions (see Theorem 4.9). We show how this lower bound results in tight sample complexity bounds for two concrete problems. The first is learning thresholds on the real line where we show a bound of $\Theta(\ln(1/\delta)/\epsilon^2)$. Then we show sample complexity of $\Theta\left(\frac{2d+\ln(1/\delta)}{\epsilon^2}\right)$ for the union of d intervals on the real line.

The sample complexity of the union of d intervals for a fixed distribution in a noisy setting has also been investigated by Gentile and Helmbold (1998). They show a lower bound of $\Omega\left(2d \log \frac{1}{\Delta}/(\Delta(1-2\eta)^2)\right)$ where Δ is the probability mass of the symmetric difference between the true target hypothesis and the output hypothesis that the algorithm should guarantee with high probability, and η is the noise parameter (i.e. the probability that a correct label is corrupted, see classification noise model of Angluin and Laird (1987)). This notation implies that the difference in true error of the target and the algorithm's output is $\epsilon = (1-2\eta)\Delta$. Setting $\eta = 1/2 - \epsilon/4$ gives $\Omega(2d/\epsilon^2)$. We note that we do not make the assumption of a constant level of noise for each unlabeled example. It turns out, however, that in our proofs we do construct worst case distributions that have a constant noise rate that is slightly below $1/2$.

We point out two main differences between our results and that of Gentile and Helmbold. The first being that we explicitly construct noisy distributions to obtain ϵ^2 in the denominator. The second difference is that our technique appears to be quite different from theirs, which uses an information theory approach, whereas we make use of known techniques based on lower bounding how well one can distinguish similar noisy distributions, and then applying an averaging argument. The main tools used in this section come from Anthony and Bartlett (1999, Chapter 5).

We first cite a result on how many examples are needed to distinguish two similar, Bernoulli distributions in Lemma 4.6. Then in Lemma 4.7 we prove an analogue of this for arbitrary unlabeled distributions. The latter result is used to give us a lower bound in Theorem 4.9 for b-shatterable distributions (see Definition 4.6). Corollary 4.8 and 4.11 gives us tight sample complexity bounds for thresholds and union of intervals on $\mathbb{R}$.

Lemma 4.6 (Anthony and Bartlett (1999))**.** *Suppose that P is a random variable uniformly distributed on $\{P_1, P_2\}$ where P_1, P_2 are Bernoulli distributions over $\{0,1\}$ with $P_1(1) = 1/2 - \gamma$ and $P_2(1) = 1/2 + \gamma$ for $0 < \gamma < 1/2$. Suppose that $\xi_1, \ldots, \xi_m$ are i.i.d. $\{0,1\}$ valued random variables with $\Pr(\xi_i = 1) = P(1)$ for each i. Let f be a function from $\{0,1\}^m \to \{P_1, P_2\}$. Then*

$$\mathop{\mathbb{E}}_{P} \Pr_{\xi \sim P^m} [f(\xi) \neq P] > \frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-4m\gamma^2}{1-4\gamma^2}\right)}\right) =: F(m, \gamma).$$

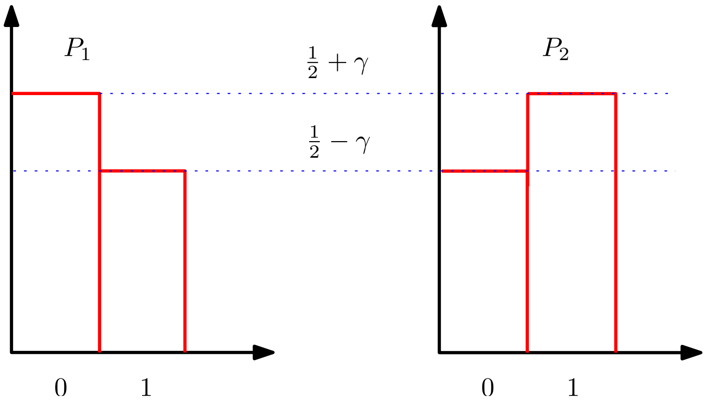

Figure 4.3: Lemma 4.6 is saying one needs at least $\Omega(\ln(1/\delta)/\gamma^2)$ examples chosen from $P \sim$ Bernoulli(P_1, P_2) to distinguish the two distributions with confidence at least δ.

See Figure 4.3 for an illustration. One can view the lemma this way: if one randomly picks two weighted coins with similar biases, then there's a lower bound on the confidence with which one can accurately predict the coin that was picked.

The next result is similar except an unlabeled distribution D is fixed, and the distributions we want to distinguish will be extensions of D.

Lemma 4.7. *Fix any $\mathcal{X}$, $\mathcal{H}$, unlabeled distribution D, and $m \geq 0$. Suppose there exists $h, g \in H$ with $D(\mathsf{set}(h)\Delta\,\mathsf{set}(g)) > 0$. Let P_h and P_g be the extension of D such that $P_h(h(x)|x) = P_g(g(x)|x) = 1/2+\gamma$. Let $\mathcal{A}_D : (h\Delta g \times \mathcal{Y})^m \to \mathcal{H}$ be any function. Then for any $x_1, \ldots, x_m \in h\Delta g$, there exists $P \in \{P_h, P_g\}$ such that if $y_i \sim P(\cdot|x_i)$ for all i,*

$$\Pr_{y_i}[\mathrm{Er}^P(\mathcal{A}_D((x_1,y_1),\ldots,(x_m,y_m))) - \mathrm{OPT}_P > \gamma D(\mathsf{set}(h)\Delta\,\mathsf{set}(g))] > F(m,\gamma),$$

where $\mathrm{OPT}_P = 1/2 - \gamma$. Thus if the probability of failure is at most δ, we require

$$m \geq \left(\frac{1}{4\gamma^2} - 1\right)\ln\frac{1}{8\delta}. \tag{4.13}$$

Proof. Suppose for a contradiction this is not true. Let $\mathfrak{P} = \{P_h, P_g\}$. Then there exists an $\mathcal{A}_D$ and $x_1, \ldots, x_m$ such that

$$\forall P \in \mathfrak{P},\ \Pr_{y_i}[\mathrm{Er}^P(\mathcal{A}_D((x_1,y_1),\ldots,(x_m,y_m))) - \mathrm{OPT}_P > \gamma D(\mathsf{set}(h)\Delta\,\mathsf{set}(g))] \leq F(m,\gamma). \tag{4.14}$$

Then we will show that the lower bound in Lemma 4.6 can be violated. Now $\mathsf{set}(h)\Delta\,\mathsf{set}(g)$ can be partitioned into

$$\Delta_0 = \{x : h(x) = 0\} \text{ and } \Delta_1 = \{x : h(x) = 1\}.$$

Without loss of generality assume $\{x_1, \ldots, x_\ell\} \subseteq \Delta_0$ and $\{x_{\ell+1}, \ldots, x_m\} \subseteq \Delta_1$. Let $\alpha = \mathcal{A}_D((x_1, y_1), \ldots, (x_m, y_m))$. From the triangle inequality

$$D(\mathsf{set}(\alpha) \Delta \,\mathsf{set}(h)) + D(\mathsf{set}(\alpha) \Delta \,\mathsf{set}(g)) \geq D(\mathsf{set}(h) \Delta \,\mathsf{set}(g)).$$

Thus if α is closer to h then $D(\mathsf{set}(\alpha) \Delta \,\mathsf{set}(g)) \geq D(\mathsf{set}(h) \Delta \,\mathsf{set}(g))/2$ and vice versa. Let P be a random variable uniformly distributed on $\mathfrak{P}$. For any setting of $P \in \mathfrak{P}$ We have

$$\Pr_{y_1 \sim P(\cdot|x_1)}(y_1 = 1) = \cdots = \Pr_{y_\ell \sim P(\cdot|x_\ell)}(y_l = 1) = P(1|x \in \Delta_0) = \\ \Pr_{y_{\ell+1} \sim P(\cdot|x_{\ell+1})}(y_{\ell+1} = 0) = \cdots = \Pr_{y_m \sim P(\cdot|x_m)}(y_m = 0) = P(0|x \in \Delta_1).$$

Let $\xi_1, \ldots, \xi_m \sim P(\cdot|x \in \Delta_0)$ so that

$$\Pr(\xi_i = 1) = \begin{cases} \frac{1}{2} - \gamma & \text{if } P = P_h \\ \frac{1}{2} + \gamma & \text{if } P = P_g \end{cases}.$$

Let us define the function $f : \{0,1\}^m \to \mathfrak{P}$ as follows. It will take as input $\xi_1, \ldots, \xi_m$ then transform this to an input of $\mathcal{A}_D$ as $I = (x_1, \xi_1), \ldots, (x_l, \xi_l), (x_{\ell+1}, 1 - \xi_{\ell+1}), \ldots, (x_m, 1 - \xi_m)$ so that ξ_i and $1 - \xi_j$ is from the same distribution as y_i and y_j, respectively, for $i \leq \ell, j > \ell$. Now define

$$f(\xi_1, \ldots, \xi_m) = \begin{cases} P_h & \text{if } D(\mathsf{set}(\mathcal{A}_D(I)) \Delta \,\mathsf{set}(h)) < D(\mathsf{set}(\mathcal{A}_D(I)) \Delta \,\mathsf{set}(g)) \\ P_g & \text{otherwise} \end{cases}.$$

We have

$$\begin{aligned} &\mathop{\mathbb{E}}_{P} \Pr_{\boldsymbol{\xi} \sim P^m(\cdot|x \in \Delta_0)} [f(\xi) \neq P] \\ &\leq \mathop{\mathbb{E}}_{P} \Pr_{\boldsymbol{\xi}} [D(\mathsf{set}(\mathcal{A}_D(I)) \Delta \,\mathsf{set}(\mathrm{OPT}_P)) > D(\mathsf{set}(h) \Delta \,\mathsf{set}(g))/2] \\ &\leq \mathop{\mathbb{E}}_{P} \Pr_{\boldsymbol{\xi}} \left[\mathrm{Er}^P(\mathcal{A}_D(I)) - \mathrm{OPT}_P > \gamma D(\mathsf{set}(h) \Delta \,\mathsf{set}(g))\right] \\ &\leq F(m, \gamma), \end{aligned}$$

where the last inequality follows from (4.14). This is a contradiction, so the lower bound from Lemma 4.6 must apply. If the probability of failure $F(m, \gamma)$ is at most δ, solving the inequality for m gives (4.13). □

Corollary 4.8. *Recall the class of thresholds is defined as $H = \{\mathbf{1}(-\infty, t] \; : \; t \in \mathbb{R}\}$. Let $\epsilon, \delta \in (0,1]^2$, and D any unlabeled distribution. Then the following holds,*

$$\min_{\mathcal{A}} \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, H, P) = \Theta\left(\frac{\ln \frac{1}{\delta}}{\epsilon^2}\right).$$

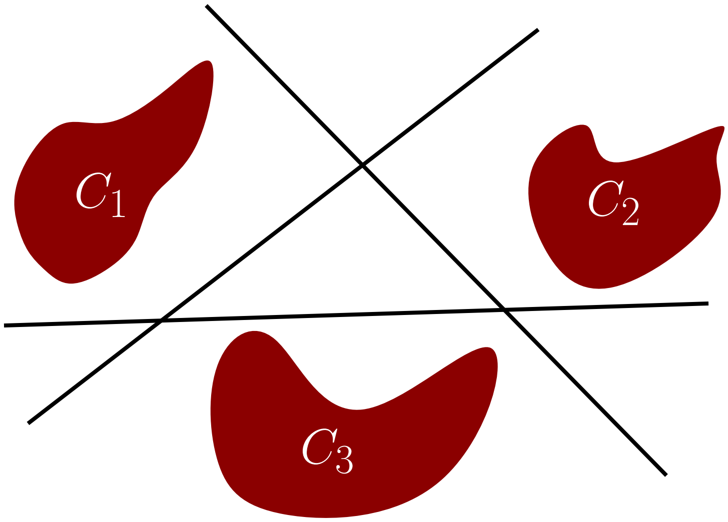

Figure 4.4: Here $\mathcal{H}$ consists of linear halfspaces over the Euclidean plane. The unlabeled distribution D has support on sets C_1, C_2 and C_3 (assigning equal probabilities to each). In this case the three sets are 3-shatterable.

Proof. Upper bound comes from any ERM algorithm (note that $\mathrm{VC}(H) = 1$, use uniform convergence results in Section 2.5). Let $h = \mathbf{1}(-\infty, 0]$ and $g = \mathbf{1}(-\infty, 1]$ so $D(\mathsf{set}(h) \Delta \mathsf{set}(g)) = 1$. Set $\gamma = \epsilon$ as in Lemma 4.7. □

Definition 4.6. Let D be an unlabeled distribution. The triple ($\mathcal{X}$, $\mathcal{H}$, D) is b-shatterable if there exists disjoint sets $C_1, C_2, \ldots, C_b$ with $D(C_i) = 1/b$ for each i, and for each $S \subseteq \{1, 2, \ldots, b\}$, there exists $h \in \mathcal{H}$ such that

$$\mathsf{set}(h) \cap \left(\bigcup_{i=1}^{b} C_i \right) = \bigcup_{i \in S} C_i.$$

See Figure 4.4 for an example.

Theorem 4.9. *If ($\mathcal{X}$, $\mathcal{H}$, D) is b-shatterable then for any $\epsilon, \delta \in (0, 1/64)^2$ we have*

$$\min_{\mathcal{A}} \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P) = \Omega\left(\frac{b + \ln \frac{1}{\delta}}{\epsilon^2} \right).$$

Proof. The proof is similar to Theorem 5.2 in Anthony and Bartlett (1999). The idea is to construct noisy distributions and use the probabilistic method to show that there is at least one noisy distribution which the *a priori* fixed SSL algorithm does badly on. Let $G = \{h_1, h_2, \ldots, h_{2^b}\}$ be the class of functions that b-shatters D with respect to $C = \{C_1, \ldots, C_b\}$. We construct noisy extensions of D, $\mathfrak{P} = \{P_1, P_2, \ldots, P_{2^b}\}$ so that for each i,

$$P_i(x, h_i(x)) = \frac{1 + 2\gamma}{2b}.$$

For any $h \in \mathcal{H}$, let

$$\mathrm{snap}(h) := \operatorname*{argmin}_{h' \in G} D(\mathsf{set}(h) \Delta \,\mathsf{set}(h')).$$

Suppose $P \in \mathfrak{P}$, let h^* denote the optimal classifier which is some $g \in G$ depending on the choice of P. If $i \neq j$ and $N(h_i, h_j)$ is the number of sets in C where h_i and h_j disagree, then $D(\mathsf{set}(h_i)\Delta \,\mathsf{set}(h_j)) \geq N(h_i, h_j)/b$, and since G is a $1/b$-packing,

$$\mathrm{Er}^P(h) \geq \mathrm{Er}^P(h^*) + \frac{\gamma}{b} N(\mathrm{snap}(h), h^*) = \frac{1}{2}\left(\mathrm{Er}^P(\mathrm{snap}(h)) + \mathrm{Er}^P(h^*)\right). \tag{4.15}$$

Modifying the proof of Anthony and Bartlett (1999, Chap. 5) with the use of Lemma 4.7 rather than Lemma 4.6 we get that there exists a $P \in \mathfrak{P}$ such that whenever $m \leq b/(320\epsilon^2)$,

$$\Pr_{S \sim P^m}\left[\mathrm{Er}^P(\mathrm{snap}(\mathcal{A}(D, S))) - \mathrm{Er}^P(h^*) > 2\epsilon\right] > \delta.$$

Whenever $\mathcal{A}$ fails, we get from (4.15)

$$\mathrm{Er}^P(\mathcal{A}(D, S)) - \mathrm{Er}^P(h^*) \geq \frac{1}{2}\left(\mathrm{Er}^P(\mathrm{snap}(h)) + \mathrm{Er}^P(h^*)\right) \geq \epsilon.$$

To get $\Omega(\ln(1/\delta)/\epsilon^2)$, we note that b-shatterability implies there exists the hypotheses h and g that predicts 0 and 1, respectively, on the support of D, now apply Lemma 4.7 with h and g. □

An easy consequence of Theorem 4.9 is that there is no semi-supervised learning algorithm that has sample complexity asymptotically better than supervised learning, taking the *worst-case* over *all* distributions P over $\mathcal{X} \times \mathcal{Y}$.

Corollary 4.10. *Fix $\mathcal{H}$, $\epsilon, \delta > 0$. Assume that $\mathcal{H}$ contains h, g with*

$$D(\mathsf{set}(h)\Delta \,\mathsf{set}(g)) = 1$$

we have the following:

$$\min_{\mathcal{A}} \sup_{P \text{ over } \mathcal{X}\times\mathcal{Y}} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P) = \Theta\left(\frac{\mathrm{VC}(\mathcal{H}) + \ln\frac{1}{\delta}}{\epsilon^2}\right)$$

Proof. Let A be a set of size $\mathrm{VC}(\mathcal{H})$ that is shattered by $\mathcal{H}$, and let D be an unlabeled distribution such that for any $a \in A$, $D(a) = 1/\mathrm{VC}(\mathcal{H})$, then we apply Theorem 4.9 since $(\mathcal{X}, \mathcal{H}, D)$ is b-shatterable. □

We will now apply Theorem 4.9 to give the SSL sample complexity for learning union of intervals on the real line, for any unlabeled distribution having density. Recall that by the reduction to the uniform distribution technique in Section 4.2, we only need to consider the sample complexity with respect to the uniform distribution on $[0, 1]$.

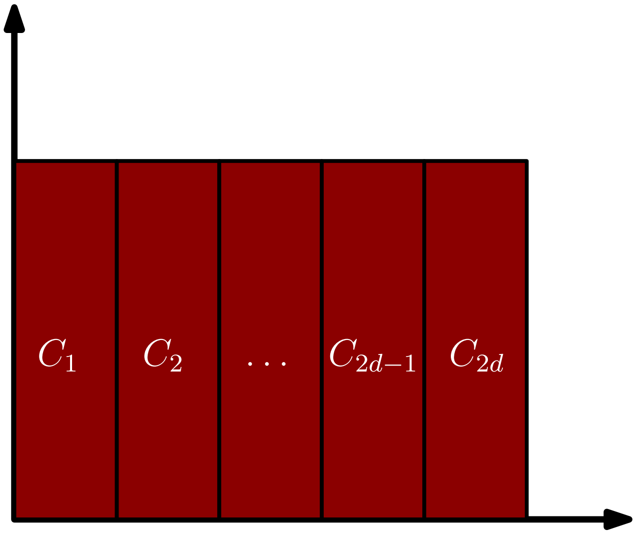

Figure 4.5: Let the unlabeled distribution D be the uniform distribution, then the sets $C_1, \dots, C_{2d}$ can be $2d$-shattered by the union of at most $2d$ intervals, UI_d.

Corollary 4.11. *Recall the class of union of at most d intervals $\mathrm{UI}_d = \{\mathbf{1}[a_1, a_2) \cup \dots \cup [a_{2l-1}, a_{2l}) : l \le d, 0 \le a_1 \le a_2 \le \dots \le a_{2l} \le 1\}$. Let $(\epsilon, \delta) \in (0,1]^2$ and D any unlabeled distribution having density. Then the following holds,*

$$\min_{\mathcal{A}} \sup_{P \in \mathsf{Ext}(D)} m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, \mathrm{UI}_d, P) = \Theta\left(\frac{2d + \ln \frac{1}{\delta}}{\epsilon^2}\right).$$

Proof. By reduction to the uniform distribution (i.e. Lemma 4.1 and Lemma 4.2) we can assume D to be uniform over $[0,1]$. We have $\mathrm{VC}(\mathrm{UI}_d) = 2d$ (see Claim A.1), thus the upper bound follows immediately, by any ERM type algorithm. For the lower bound, construct $2d$-shatterable sets by letting

$$C_i = \left[\frac{i-1}{2d}, \frac{i}{2d}\right)$$

for $i = 1, \dots, 2d$. For any $S \subseteq \{1, \dots, 2d\}$ define $h_S = \mathbf{1} \bigcup_{i \in S} C_i$. Now if $|S| \le d$ then clearly $h_S \in \mathrm{UI}_d$, if $|S| > d$ then $h_{\overline{S}} \in \mathrm{UI}_d$ since $|\overline{S}| < d$. But then $[0,1) \backslash h_{\overline{S}}$ can be covered by at most d intervals, so $h_S \in \mathrm{UI}_d$. Thus the set $\{h_S : S \subseteq \{1, \dots, 2d\}\}$ $2d$-shatters D on $[0,1]$. Also let $h = [0,0) = \emptyset$ and $g = [0,1)$. Now apply Theorem 4.9 for the bound. See Figure 4.5. □

4.5 No Optimal Semi-Supervised Algorithm

One could imagine a different formulation of the comparison between supervised learning and semi-supervised learning. For example, one might ask naively whether, for given class $\mathcal{H}$, there is a semi-supervised algorithm $\mathcal{A}$, such that for any supervised algorithm $\mathcal{B}$, and any sufficiently small ϵ, δ, on any data generating distribution P, the sample complexity of $\mathcal{A}$ is no larger than the sample complexity of $\mathcal{B}$. The answer to the question is easily seen to be negative, because for any P there exists a supervised learning algorithm B_P that ignores the labeled examples and simply outputs the hypothesis $h \in \mathcal{H}$ with minimum error $\mathrm{Er}^P(h)$ (or even the Bayesian optimal classifier for P). On P the sample complexity of B_P is zero, unfortunately, on P', sufficiently different from P, the sample complexity of B_P can be infinite.

One might disregard algorithms such as B_P and ask the same question as above, except that one quantifies over only the subset of algorithms that on *any* distribution over $X \times \{0, 1\}$ have sample complexity that is distribution-*free* (e.g. polynomial in $1/\epsilon$ and $\ln(1/\delta)$). That is, restricting oneself to learning algorithms (see Definition 2.5). The following theorem demonstrates that such restriction does not help and the answer to the question is still negative.

Theorem 4.12. *Let $H = \{\mathbf{1}(-\infty, t] \ : \ t \in \mathbb{R}\}$ be the class of thresholds over the real line. For any unlabeled distribution D having density (with respect to the Lebesgue measure), any semi-supervised algorithm $\mathcal{A}$, any $\epsilon > 0$ and $\delta \in (0, \frac{1}{2})$, there exists a distribution $P \in \mathsf{Ext}(D)$ and a supervised learning algorithm $\mathcal{B}$ (see Definition 2.8) such that*

$$m_{\mathrm{SSL}}(\mathcal{A}, \epsilon, \delta, H, P) > m_{\mathrm{SL}}(\mathcal{B}, \epsilon, \delta, H, P) \ .$$

Proof. Fix any $\mathcal{A}$, D unlabeled distribution having density, and m. Let $\mathsf{LeftERM}$ be the algorithm that chooses the right most positive point (for precise definition, see Equation 2.3). For any $h \in H$ we also define algorithm

$$\mathsf{LeftERM}_h(S) := \begin{cases} h & \text{if } \mathrm{Er}^S(h) = 0 \\ \mathsf{LeftERM}(S) & \text{otherwise.} \end{cases}$$

First, note that $\mathsf{LeftERM} = \mathsf{LeftERM}_{\mathbf{1}\emptyset}$. Second, for any h, $\mathsf{LeftERM}_h$ is an ERM paradigm (see Definition 2.11), and since $\mathrm{VC}(H) = 1$, it is a learning algorithm. Third, clearly the sample complexity of $\mathsf{LeftERM}_h$ on D_h is zero independently of the choice of ϵ and δ.

Theorem 4.5 shows that there exists a $h \in H$ such that the sample complexity of $\mathcal{A}$ on D_h is positive, in fact, it is increasing as ϵ and δ approach zero. Thus, there exists a supervised algorithm $\mathcal{B} = \mathsf{LeftERM}_h$ with lower sample complexity than $\mathcal{A}$. □

Chapter 5

Conclusion

This thesis has presented a fresh, new look at the theoretical underpinnings of semi-supervised learning. We proposed a novel, utopian model of semi-supervised learning where we assume the entire unlabeled distribution is given to the learner and thus bypassing sampling issues of unlabeled training data.

Our analysis is concerned with the inherent limitations of semi-supervised learning. In particular, we analyze the potential sample complexity gains in our utopian model of semi-supervised learning compared with that of supervised learning. All these analyses are done under what can be termed as the *no-prior-knowledge* setting, where no assumptions are made on the relationship between labels and the unlabeled data distribution.

> *Our main point is that semi-supervised learning cannot provide significant benefits over supervised learning, unless one is absolutely sure that an assumption holds on the relationship between labels and the unlabeled data distribution.*

More technically, what we have proved is that for the class of thresholds (realizable and agnostic) and union of intervals (agnostic) SSL cannot have better than a constant factor advantage in the sample complexity. We conjecture this to be a more general phenomenon that applies to *any* hypothesis class. Of course, this does not take into account the situation in which the learner makes SSL type assumptions (e.g. the cluster assumption) *and* is positive these assumptions hold. However, the difficulty of verifying SSL assumptions or mathematically formalizing them is still a wide open question. For example, we have illustrated counter-intuitive hazards of learning under the cluster assumption.

Our work is the first in addressing the fundamental limitations of semi-supervised learning in a theoretical setting. We hope it is also a first step in bridging the gap between theory and practice of semi-supervised learning. There is still much future work to be done and a few of them are given below.

5.1 Proving Conjectures 4.1 and 4.2

The main reason why one should believe the conjecture to be true (at least for the realizable case) is the algorithm of Kääriäinen (2005). That is, a generalization of MedianERM for arbitrary domains $\mathcal{X}$ where one finds the "centre" of the version space (i.e. consistent hypothesis on labeled data) where the (pseudo)-metric over the hypothesis space is defined by the unlabeled distribution mass of the symmetric difference of two hypotheses. Intuitively, supervised learning cannot find this centre, therefore in the worst-case it chooses a hypothesis on the "boundary" of the version space resulting in error at most the diameter of the version space whereas SSL achieves error at most the radius of the version space.

While we have shown the conjecture to be true for "smooth" unlabeled distributions (i.e. ones having a density function) for the class of thresholds in both the realizable and agnostic settings, and union of intervals in the agnostic setting, it would be nice to prove it for unlabeled distributions that don't have densities. These distributions include ones that are discrete, or a combination of discrete and continuous parts. For example, if the distribution is concentrated on one point, the best SSL and SL algorithms do the same thing.

It would also be nice to prove a lower bound for supervised learning under thresholds in the realizable setting (for unlabeled distributions having density). We believe the lower bound should match the best known upper bound (e.g. LeftERM) of $\ln(1/\delta)/\epsilon$ within some negligible additive factor. This would imply that for any unlabeled distribution having density, the SSL sample complexity is *exactly* twice better than SL sample complexity—without making SSL assumptions!

Of course, it will be quite interesting to prove the conjecture for classes in higher dimensions. Unfortunately the reduction technique of Section 4.2 cannot be easily extended for higher dimensions. For example, consider $\mathbb{R}^2$, and the map $(x, y) \mapsto (F_1(x), F_2(y))$ where F_1 and F_2 are the CDFs for the x and y components, respectively. In order for this mapping to induce a uniform distribution over $[0, 1]^2$, we need that $F(x, y) = F_1(x)F_2(y)$ where F is the CDF over $\mathbb{R}^2$. In other words F must be a product distribution (having density). The class of axis-aligned rectangles will retain its "shape" under this type of transformation, but not necessarily for linear halfspaces (e.g. product Gaussian truncated beyond the unit square and halfspace defined by $y = 1 - x$). However, we do not know of a lower bound on learning axis-aligned rectangles with respect to the fixed uniform distribution over the unit cube. Note that there are still transformations that will induce the uniform distribution for non-product distributions, but it may disfigure the shape of say, axis-aligned rectangles.

5.2 Verifying SSL Assumptions

Our examples in Section 3.3 underscores the pitfalls when doing SSL with the cluster assumption. A natural question to consider is: is there's a way to verify these SSL assumptions?

Our examples show that the cluster assumption hold at varying degrees. That is, while the best separator may not lie in the *least dense region* of the unlabeled distribution, it may lie at some region with density slightly larger. Thus when verifying this assumption it is important to *verify at what degree the assumption holds.*

The main issue is that in order to verify the degree to which a SSL assumption holds, one will likely need to set aside existing labeled data to do so. After performing the verification, SSL will have less labeled data to train with. However, supervised learning does not need perform this kind of verification, therefore it has more labeled training data, which can result in a predictor that is just as good as the SSL one.

Perhaps a way to overcome this issue is to consider the setting when one has related tasks in the sense that they share the same labels/unlabeled distribution relationship. In practice a learner may use SSL assumptions for the current task that are known to work for related tasks. The main potential advantage here is that once the assumption has been verified on another task, it can be applied to the related task without using more labeled data for further verification. The definition of "relatedness" is important here, and the learner's new problem is to make sure the tasks are related.

5.3 SSL Sample Complexity Incorporating Assumptions

As discussed in Section 4.1, the analysis of the SSL sample complexity does not consider SSL assumptions and is a worst-case analysis. Specifically, for a fixed hypothesis class $\mathcal{H}$, unlabeled distribution D, $\epsilon, \delta > 0$,

$$\text{Worst-case-}m_{\text{SSL}}(\epsilon, \delta, \mathcal{H}, D) := \min_{\mathcal{A}} \sup_{P \in \mathsf{Ext}(D)} m_{\text{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P)$$

The supremum is over $P \in \mathsf{Ext}(D)$, that is, worst-case (labeled) distribution whose marginal equals D. This is somewhat unrealistic, as it allows for P's that is very noisy, which would result in bad performance for both SSL and SL.

How can this worst-case be modified to be more realistic for SSL with assumptions? A possible idea is to consider a restricted subset of $\mathsf{Ext}(D)$ that satisfy an assumption. That is, let T be an assumption (e.g. cluster assumption) and let $T(D) \subseteq \mathsf{Ext}(D)$ such that each $P \in T(D)$ satisfy the assumption T. Thus we can define an *assumption*-based SSL sample complexity,

$$\text{Assumption-}T\text{-}m_{\text{SSL}}(\epsilon, \delta, \mathcal{H}, D) := \min_{\mathcal{A}} \sup_{P \in T(D)} m_{\text{SSL}}(\mathcal{A}, \epsilon, \delta, \mathcal{H}, P)$$

In the extreme case $T(D)$ consists of only a single distribution and the (labeled) sample complexity is zero. For example, if $T(D)$ consists of the distribution that is defined by the lowest density linear separator (of D), then the task is to find this separator using only unlabeled data, since we know it has zero error. Clearly this is too strong an assumption. On the other extreme, T consists of all $\mathsf{Ext}(D)$ which probably will not provide a significant advantage over supervised learning. Thus it is up to the learner to specify a T that takes into account this tradeoff.

APPENDICES

Appendix A

Examples of VC Dimension

This appendix shows some examples of the VC dimension of some basic hypothesis classes. These classes include thresholds over the real line, finite union of intervals and linear halfspaces in Euclidean space. In this thesis we make use of the VC dimension of the first two hypothesis classes, but the calculations for halfspaces is left for didactical purposes.

Recall from Definition 2.10 that the Vapnik-Chervonenkis dimension of a hypothesis class $\mathcal{H}$ over input domain $\mathcal{X}$ is defined as

$$\mathrm{VC}(\mathcal{H}) := \sup\{d : \exists A \subseteq \mathcal{X},\ |A| = d,\ |\mathcal{H}(A)| = 2^d\},$$

where $\mathcal{H}(A) = \{h(A) : h \in \mathcal{H}\}$. That is, it is the size of the largest subset A of $\mathcal{X}$ such that for any $S \subseteq A$, there is a $h_S \in \mathcal{H}$ with $\mathsf{set}(h_S) \cap A = S$.

A.1 Thresholds

The class of *thresholds* over the real line is defined as

$$H = \{\mathbf{1}(-\infty, a] : a \in \mathbb{R}\}.$$

Figure A.1 shows an example hypothesis.

It is not hard to see that H can shatter one point, say 0, by the hypothesis $\mathbf{1}(-\infty, -1]$ and $\mathbf{1}(-\infty, 1]$. Let $a, b \in \mathbb{R}$ be any two points. Does there exist a t such that $(-\infty, t] \cap \{a, b\} = \{b\}$? No, because then $t \geq b$ which implies a must be in the intersection. Thus, $\mathrm{VC}(H) = 1$.

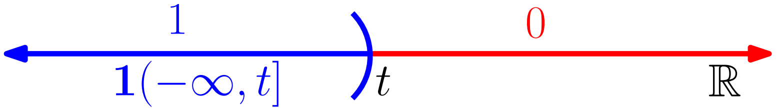

Figure A.1: The class of thresholds consists of hypothesis that predicts 1 if a point is less than the threshold, otherwise it predicts 0.

A.2 Union of Intervals

We define the union of at most d intervals as

$$\mathrm{UI}_d := \{\mathbf{1}[a_1, a_2) \cup [a_3, a_4) \cup \cdots \cup [a_{2\ell-1}, a_{2\ell}) \ : \ \ell \le d, \ a_1 \le a_2 \le \cdots \le a_{2\ell}\}.$$

Claim A.1. $\mathrm{VC}(\mathrm{UI}_d) = 2d$.

Proof. For the lower bound, let $A = \{\frac{1}{2d} : 0 \le i \le 2d-1\}$. For any subset S of A, suppose $|S| \le d$, then clearly we can construct

$$h_S = \mathbf{1}\bigcup_{i=1}^{|S|}\left[s_i, s_i + \frac{1}{2d}\right)$$

where s_i's are the elements of S, such that $\mathsf{set}(h_S) \cap A = S$. If $|S| > d$ then $|A \backslash S| < d$, this implies there exists $h_{A\backslash S}$ such that its set intersection with A is $A\backslash S$. But note that $\overline{h_{A\backslash S}}$ is a union of at most d intervals—technically not true, but can be easily modified into one—whose set intersection with A results in S.

For the upper bound, let A consist of (at least) $2d+1$ points $a_1, a_2, \ldots, a_{2d+1}$, it can be seen that $S = \{a_1, a_3, \ldots, a_{2d-1}, a_{2d+1}\}$ of size $d+1$ requires $d+1$ disjoint intervals to cut away S from A. □

A.3 Linear Halfspaces

The hypothesis class of the linear halfspaces in Euclidean dimension d is defined by

$$\mathfrak{L}_d = \{I(w_1x_1 + w_2x_2 + \ldots + w_dx_d + w_0 \ge 0) \ : \ \mathbf{w} \in \mathbb{R}^{d+1}, \mathbf{x} \in \mathbb{R}^d\},$$

where $I(\cdot)$ is the indicator function.

Claim A.2. $\mathrm{VC}(\mathfrak{L}_d) \ge d+1$.

Proof. Let $A = \{\mathbf{0}, e_1, e_2, \ldots, e_d\} \subset \mathbb{R}^d$ where e_i is the vector with all zeroes except at the i-th coordinate where it is a 1. Let

$$S' = \{e_{\ell_1}, \ldots, e_{\ell_k} : \ell_1 \le \ell_2 \le \cdots \le \ell_k\} \subset A.$$

We let $w_0 = -1/2$, $w_i = 1$ if $i = \ell_j$ for some $j \in [k]$, and $w_i = 0$ for all other i's. We let $h_{S'}$ be the halfspace defined by the given weights. It is clear that $\mathsf{set}(h_{S'}) \cap A = S'$. Also if $S = S' \cup \{\mathbf{0}\}$ then we adjust $w_0 = 0$ to include the zero vector. This shows that we can "cut out" all the possible subsets of A from A. □

In turns out this is the largest sized set that is shatterable by halfspaces (equivalent to number of parameters of the halfspace). To show the upper bound, we need a simple observation.

Definition A.1. The *convex hull* of a *finite*[1] set of points $S = \{s_1, \ldots, s_m\} \subseteq \mathbb{R}^n$ is

$$\mathsf{CH}(S) = \left\{ \lambda_1 s_1 + \ldots + \lambda_m s_m : \boldsymbol{\lambda} \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \right\}.$$

Theorem A.3 (Radon)**.** *Let $A = \{a_1, \ldots, a_{d+2}\} \subset \mathbb{R}^d$, then there exists two partitions of A, P, Q such that $\mathsf{CH}(P) \cap \mathsf{CH}(Q) \neq \emptyset$.*

Proof. Consider the vectors $A' = \{(a_1, 1), \ldots, (a_{d+2}, 1)\} \subset \mathbb{R}^{d+1}$ where an extra 1 is added to the last coordinate. Basic linear algebra says that A' is linearly dependent, that is, there exists constants $c_1, \ldots, c_{d+2} \in \mathbb{R}$ not all zero such that

$$c_1(a_1, 1) + c_2(a_2, 1) + \ldots + c_{d+2}(a_{d+2}, 1) = 0.$$

This implies these separate conditions,

$$\begin{aligned} \sum_{i=1}^{d+2} c_i a_i &= 0 \\ \sum_{i=1}^{d+2} c_i &= 0 \end{aligned}$$

Let $P = \{a_i : c_i > 0\}$ and $Q = A \backslash P$. By the above equations, the point

$$\sum_{i:a_i \in P} \left[\frac{c_i}{\sum_{j:a_j \in P} c_j} \right] a_i = \sum_{i:a_i \in Q} \left[\frac{c_i}{\sum_{j:a_j \in Q} c_j} \right] a_i$$

is in both $\mathsf{CH}(P)$ and $\mathsf{CH}(Q)$. □

Corollary A.4. $\mathrm{VC}(\mathfrak{L}_d) = d + 1$.

Proof. The upper bound comes from Claim A.2. Given any A with at least $d + 2$ points, we can find a partition of A, say P, Q such that their convex hulls do not intersect. A basic result in convex analysis says that a sufficient and necessary condition for linear separability of two sets of points is that their convex hulls do not intersect. This means P and Q cannot be linearly separated, therefore neither P nor Q can be "cut out" from A. □

[1] The convex hull of any set (e.g. uncountable) can also be defined, but we will not need it here.

References

Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1987. 49

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, January 1999. 6, 16, 17, 49, 52, 53

Maria-Florina Balcan and Avrim Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of 18th Annual Conference on Learning Theory*, pages 111–126. Springer, 2005. 4, 22, 26

Maria-Florina Balcan and Avrim Blum. An augmented PAC model for semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 21, pages 61–89. MIT Press, September 2006. 4, 22, 24

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Conference on Learning Theory*, pages 33–44, 2008. 4, 37

Shai Ben-David, Tyler Lu, David Pal, and Miroslava Sotakova. Learning low density separators. In *Proceedings of Artificial Intelligence and Statistics*, 2009. To appear. 33

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991. 22, 24, 25, 31

Patrick Billingsley. *Probability and Measure*. Wiley-Interscience, 3 edition, 1995. 6

Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled sata with co-training. In *COLT*, pages 92–100, 1998. 23

Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. In *COLT*, pages 9–18, 1988.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. 8, 16

R. Bruce. Semi-supervised learning using prior probabilities and em. In *IJCAI Workshop on Text Learning: Beyond Supervision*, August 2001. 23

Vittorio Castelli. *The relative value of labeled and unlabeled samples in pattern recognition*. PhD thesis, Stanford University, Stanford, CA, December 1994. 23

Vittorio Castelli and Thomas M. Cover. 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, 1996. 23

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, September 2006. 21, 23, 25, 26

Fabio Cozman and Ira Cohen. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 4, pages 57–72. MIT Press, September 2006. 23

Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. Pac generalization bounds for co-training. In *NIPS*, pages 375–382, 2001. 23

Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability)*. Springer, February 1997. ISBN 0387946187. 6

Andrzej Ehrenfeucht, David Haussler, Michael J. Kearns, and Leslie G. Valiant. A general lower bound on the number of examples needed for learning. *Inf. Comput.*, 82(3):247–261, 1989. 16

Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *COLT*, pages 35–49, 2006. 22

Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In *COLT*, pages 157–171, 2007. 21

Claudio Gentile and David P. Helmbold. Improved lower bounds for learning from noisy examples: and information-theoretic approach. In *Proceedings of COLT 1998*, pages 104–115. ACM, 1998. 49

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. 6, 18

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. 14

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999. 23, 26

Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Proceedings of COLT 2005*, pages 127–142. Springer, 2005. vii, 4, 22, 25, 30, 40, 41, 57

Michael J. Kearns and Umesh V. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, USA, 1994. ISBN 0-262-11193-4. 6, 42

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 341–352, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi: http://doi.acm.org/10.1145/130385.130424. 6, 18

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994. 6, 18

Philip M. Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

Philip M. Long. An upper bound on the sample complexity of PAC-learning half-spaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

Joel Ratsaby and Santosh S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *COLT*, pages 412–417, 1995. 23

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007. 27

A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *NIPS*, 2008.

F. Oles T. Zhang. A probability analysis on the value of unlabeled data for classification problems. In *ICML*, pages 1191–1198, 2000. 23

M. Talagrand. Sharper bounds for gaussian and empirical processes. *Annals of Probability*, 22(28-76), 1994. 14

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984. 3, 4, 6, 18

Vladimir Vapnik and Alexei Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. *Theoretical Probability and Its Applications*, 2: 264–280, 1971. 4, 6, 13, 14

Vladimir Vapnik and Alexei Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. in Russian. 16

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998. 6

Vladimir N. Vapnik. Transductive inference and semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 24, pages 453–472. MIT Press, September 2006. 4, 21

Xiaojin Zhu. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, 2005.

Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Science, University of Wisconsin Madison, 2008. 2