Theory of Active Learning

Steve Hanneke
steve.hanneke@gmail.com

Version 1.1
September 22, 2014

An abbreviated version of this article appears in the *Foundations and Trends in Machine Learning* series [Hanneke, 2014]. This article will be updated from time to time as this field continues to develop. The latest version is available from the author's website (presently `http://www.stevehanneke.com`).

Abstract

Active learning is a protocol for supervised machine learning, in which a learning algorithm sequentially requests the labels of selected data points from a large pool of unlabeled data. This contrasts with passive learning, where the labeled data are taken at random. The objective in active learning is to produce a highly-accurate classifier, ideally using fewer labels than the number of random labeled data sufficient for passive learning to achieve the same. This article describes recent advances in our understanding of the theoretical benefits of active learning, and implications for the design of effective active learning algorithms. Much of the article focuses on a particular technique, namely disagreement-based active learning, which by now has amassed a mature and coherent literature. It also briefly surveys several alternative approaches from the literature. The emphasis is on theorems regarding the performance of a few general algorithms, including rigorous proofs where appropriate. However, the presentation is intended to be pedagogical, focusing on results that illustrate fundamental ideas, rather than obtaining the strongest or most general known theorems. The intended audience includes researchers and advanced graduate students in machine learning and statistics, interested in gaining a deeper understanding of the recent and ongoing developments in the theory of active learning.

Contents

1

Introduction

Active learning is a general protocol for supervised machine learning, involving interaction with an expert or oracle. Though there are many variants of active learning in the literature, the focus of this article is the so-called *pool-based* active learning model. Specifically, we suppose the user has obtained a (typically large) number of unlabeled data points (i.e., only the features, or covariates, are present), referred to as the unlabeled *pool*. The learning algorithm is permitted complete access to these unlabeled data. It additionally has access to an expert or oracle, capable of providing a label for any instance in this pool upon request, where the label corresponds to the concept to be learned. The queries to this expert can be sequential, in the sense that the algorithm can observe the responses (labels) to its previous requests before selecting the next instance in the pool to be labeled. As is typically the case in supervised machine learning, the objective is to produce a classifier such that, if presented with fresh unlabeled data points from the same data source, the classifier would typically agree with the label the expert would produce if he or she were (hypothetically) asked. We are especially interested in algorithms that can achieve this objective without requesting too many labels from the expert. In this regard, the active learning protocol enables us to design more powerful learning methods compared to the traditional model of supervised learning (including semi-supervised learning), here referred to as *passive learning*, in which the data points to be labeled by the expert are effectively selected at random from the pool. Indeed, the driving question in the study of active learning is how many fewer labels are sufficient for an active learning algorithm to achieve a given accuracy, compared to the number of labels necessary for a passive learning algorithm to achieve the same.

The motivation for active learning is that, in many machine learning problems, unlabeled data are quite inexpensive to obtain in abundance, while labels require a more time-consuming or resource-intensive effort to obtain. For instance, consider the problem of webpage classification: say, classifying a webpage as being about "news" or not. A basic web crawler can very quickly collect millions of web pages, which can serve as the unlabeled pool for this learning problem. In contrast, obtaining labels typically requires a human to read the text on these pages to determine whether it is a news article or not. Thus, the time-bottleneck in the data-gathering process is the time spent by the human labeler. It is therefore desirable to minimize the number of labels required to obtain an accurate classifier. Active learning is a natural approach to doing so, since we might hope to reduce the amount of redundancy in the labels provided by the expert by only asking for labels that we expect to be, in some sense, quite informative, given the labels already provided up to that time.

1.1 Why Do We Need a Theory of Active Learning?

The potential for active learning to achieve accuracies comparable to passive learning using fewer labels has been observed in many practical applications over the past several decades. However, intermixed with these shining positive outcomes has been an equally-vast array of applications for which these same active learning methods failed to provide any benefits; some of these algorithms have even been observed to perform *worse* than their passive learning counterparts in certain application domains. How should we interpret these negative outcomes? Is the active learning protocol fundamentally unable to provide any benefits in these application domains, or might these observations simply reflect the need to develop smarter active learning algorithms? Questions such as these beg for a theoretical treatment. More abstractly, we are asking what kind of performance we should expect from a well-designed active learning algorithm, so that we may evaluate whether a given method meets this standard. Is it reasonable to expect an algorithm to always provide improvements over passive learning, or will there be some applications where no active learning strategy can outperform a given passive learning strategy? In the scenarios where active learning is potentially beneficial, how many fewer labels should we expect a well-designed active learning algorithm to require for obtaining a given accuracy? Attempts to answer these questions naturally lead us to a deeper understanding of the general principles that should underly well-designed active learning algorithms, so that the result of such an investigation is both a better understanding of the fundamental capabilities of active learning, and insights that can guide the design of practical active learning algorithms.

A second motivation for developing a theory of active learning is that, as will hopefully be apparent in the presentation below, many wonderfully beautiful and elegant mathematical concepts and theorems arise quite naturally out of the active learning formalism. We are incredibly lucky that such a natural framework for interactive machine learning can be studied in such generality, with many general properties concisely characterized by such simple mathematical constructions. For reasons such as these, the exploration of this fascinating mathematical landscape has become a source of satisfaction and joy for many in the growing community of active learning researchers.

1.2 What is Covered in This Article?

This article includes some of the recent advances in the theory of active learning, focusing on characterizing the number of label requests sufficient for an active learning algorithm to achieve a given accuracy;

this number is known as the *label complexity*. As our interest in active learning is in its ability to reduce the label complexity compared to passive learning, we will also review some of the known results for passive learning, so as to establish a baseline for comparison.

Throughout much of the article, we will focus on one particular active learning technique, known as *disagreement-based* active learning. The reason for this choice is that the literature on disagreement-based active learning represents a fairly coherent, elegant, and mature thread in the broader active learning literature, and is now quite well-understood, with a rich variety of established results. It provides us a unified approach to active learning, which can be applied with essentially any classifier representation, can be studied under a variety of noise models, and composes well with standard relaxations that enable computational efficiency (namely, the use of surrogate losses). The established results bounding the label complexity of this technique are concise, easy to comprehend, and often fairly tight (in the sense that the algorithm actually requires nearly that many labels).

However, it is known that disagreement-based active learning is sometimes not optimal. For this reason, we additionally discuss several alternative techniques, most of which are more involved and less understood, but which are known to sometimes yield smaller label complexities than disagreement-based methods. As the literature on these other techniques is less developed, our discussion of each of them will necessarily be somewhat brief; however, some of these approaches represent important directions for investigation, and further development of these techniques would undoubtedly be of great value.

The basic outline of the article is as follows. Chapter 2 introduces the formal setting, some basic notation, and essential definitions, along with a few basic examples illustrating the fundamental concepts, style of analysis, and typical results. Chapter 3 briefly surveys the known results on the label complexity of passive learning, which serve as a baseline for comparison throughout. Chapter 4 describes several known lower bounds on the label complexity of active learning, which provide an additional point of comparison, particularly in discussions of optimality. Chapter 5 introduces the basic idea of disagreement-based active learning, along with a thorough analysis of the technique for the simple scenario of noise-free learning (the so-called *realizable case*). This is followed by a description of a noise-robust variant of the disagreement-based learning strategy, and an analysis of its label complexity under various commonly-studied noise conditions. In Chapter 6, we discuss a simple trick, involving the use of a convex relaxation of the loss function, which can make the previously-discussed algorithm computationally efficient, while still allowing us to provide formal guarantees on its label complexity under certain restricted conditions. The results concerning the label complexity of disagreement-based active learning are expressed in terms of a simple quantity, known as the *disagreement coefficient*. Chapter 7 is dedicated to describing the known properties of the disagreement coefficient, including sufficient conditions for it to obtain favorable values, and several specific learning problems for which the value of the disagreement coefficient has been calculated. Finally, Chapter 8 briefly surveys several of the other threads from the literature on the theory of active learning. It is worth mentioning that the dependences among several of these chapters are rather weak. In particular, most of the discussion of bounds on the disagreement coefficient in Chapter 7 can be read anytime after Chapter 2. Additionally, the discussion of surrogate losses in Chapter 6 can be considered largely optional in the sequence, and may be skipped without significant loss of continuity (aside from dependences in Section 8.8).

Much of the article is structured around a few algorithms, emphasizing several theorems concerning their respective label complexities, along with a variety of results on the relevant quantities those results are expressed in terms of. Where appropriate, I have accompanied these results with rigorous proofs. However, as this discussion is intended to be pedagogical, in many cases I have refrained from presenting the strongest or most general form of the results from the literature, instead choosing a form that clearly illustrates the fundamental ideas without requiring too many additional complications; the article includes numerous references to the literature where the interested reader can find the stronger or more general forms of the results. I have also attempted to provide high-level reasoning for each of the main results, so that casual readers can grasp the core ideas motivating the algorithms and leading to the formal theorems, without needing to wade through the details needed to convert the ideas into a formal proof. The technical content of this article is intended to be suitable for researchers and advanced graduate students in statistics or machine learning, familiar with the basics of probability theory and statistical learning theory at the level of an introductory graduate course.

1.3 Conceptual Themes

Before beginning the technical discussion, we first briefly illustrate some of the main concepts that arise below. Readers completely unfamiliar with active learning may also find the brief survey of Dasgupta [2011] helpful, as it provides a concise and lucid description of the main themes, without getting into as much technical detail as the present article.

As mentioned, the focus of much of this article is on the strategy of *disagreement-based* active learning, an elegant and general idea introduced in the seminal work of Cohn, Atlas, and Ladner [1994]. To illustrate this idea, consider the problem of learning a *linear separator* in the 2-dimensional plane: that is, the label of each point is "+" if the point is on one side of a particular (unknown) line, called the *target separator*, and is "$-$" if the point is on the other side. Suppose, at some time, we have observed a few labeled data points, as in Figure 1.1a. We know the target separator is some line that separates all of the "+" points from the "$-$" points; a few such lines are depicted in Figure 1.1b (in truth, there are an infinite number of possibilities). If we are then given a new unlabeled point, such as the one marked "$\circ$" in Figure 1.1c, the question is whether or not we should request its label. In this particular case, note that *all* of the lines separating the observed "+" points from the observed "$-$" points have this new point on the "$-$" side of the line. Since we know the target separator is among these lines, we can conclude that the correct label of this new point is "$-$". The important detail here is that *we did not need to observe the correct label in order to deduce its value.*

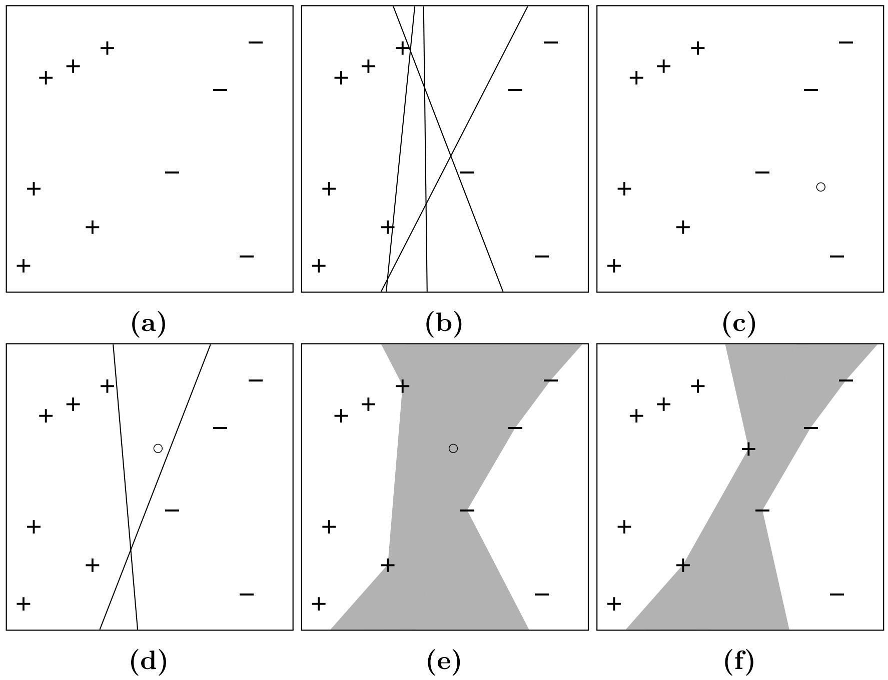

Figure 1.1: An illustration of the concepts involved in disagreement-based active learning, in the context of learning a linear separator in 2 dimensions.

On the other hand, what if instead we are given the unlabeled point depicted in Figure 1.1d? In this case, there is some line that correctly separates the other points while including this new point on the "−" side, and there is another line that correctly separates the other points while including this new point on the "+" side. So we are unable to deduce the correct label of this point based only on the information already available. The disagreement-based active learning strategy is characterized by the fact that it will *request* the value of the label (from the expert/oracle) whenever (and only whenever) this is the case. Indeed, for this data set, the disagreement-based strategy would make a label request when presented with any unlabeled point in the shaded region in Figure 1.1e: namely, the set of points such that there is some disagreement among the separators consistent with the observed labels. This set is referred to as the *region of disagreement* (or region of uncertainty).

Since the disagreement-based active learning strategy requests the label of a sample only if it is in the region of disagreement, the analysis of the label complexity of this strategy hinges on understanding the probability a new sample will be inside the region of disagreement. In particular, we will be interested in how this probability behaves as a function of the number of observed data points. The good news is that often (though not always) this probability decreases as the data set grows. For instance, suppose, in response to our request, we are told that the label of the new point in Figure 1.1d is "+". If we then add this point to the data set, the *new* region of disagreement becomes the shaded region in Figure 1.1f, which is a significant reduction compared to the region in Figure 1.1e (e.g., under a uniform probability measure within the figure). In the next chapter, we will introduce a quantity called the *disagreement coefficient*, which helps us to characterize the *rate of decrease* of the probability of getting a point in the region of disagreement.

One of the most remarkable facts about this idea is that it is fully *general*, in the sense that the exact same principle can be used in combination with *any* type of classifier. For instance, consider instead the problem of learning an *axis-aligned rectangle* in the 2-dimensional plane: that is, the label of each point is "+" if the point is contained inside an (unknown) rectangle $[a_1, b_1] \times [a_2, b_2]$ in the plane, and is "$-$" if the point is outside this rectangle. Suppose we have obtained a data set as depicted in Figure 1.2a. A few of the rectangles consistent with these labels are depicted in Figure 1.2b (again, there are in fact an infinite number of consistent rectangles). The region of disagreement is then depicted as the shaded region in Figure 1.2c. Thus, if we are given a new sample outside this shaded region, we can deduce its label without requesting its value; in the interior unshaded region, the deduced label would be "+", while in the exterior unshaded region, the deduced label would be "$-$". Again, the disagreement-based active learning strategy would request the label of a new point if and only if it is inside the shaded region. As before, given the requested label of a point in the shaded region, adding this labeled point to the data set

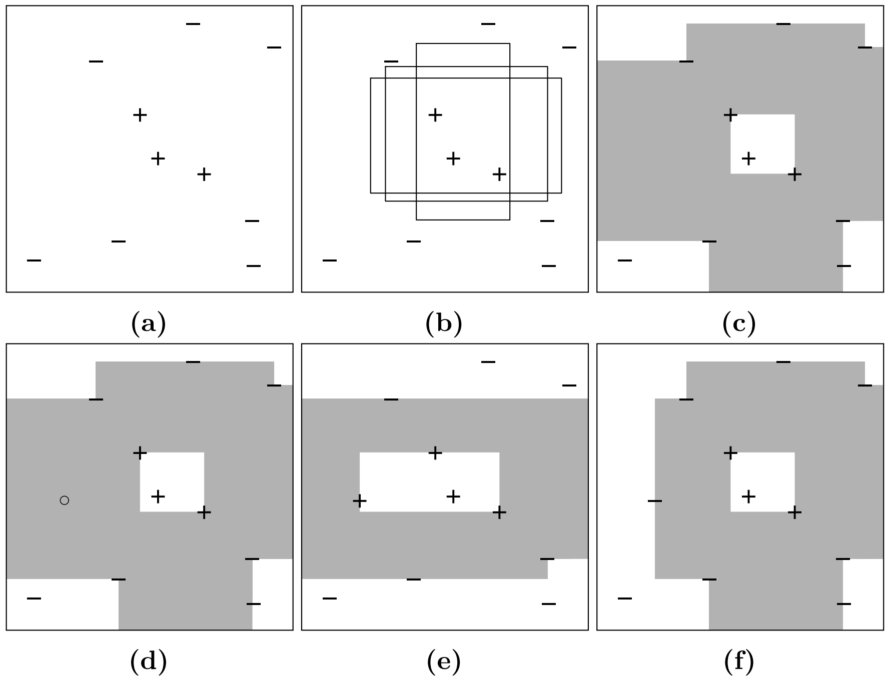

Figure 1.2: The same core idea of disagreement-based active learning can be applied with any type of classifier. Here we illustrate these concepts in the context of learning an axis-aligned rectangle in 2 dimensions.

would cause a reduction in the region of disagreement. For instance, for the new point marked "$\circ$" in Figure 1.2d, if we are told the correct label is "+", upon adding this point to the data set, the new region of disagreement would be the shaded region depicted in Figure 1.2e; on the other hand, if we are told the correct label is "$-$", the new region of disagreement would be the shaded region depicted in Figure 1.2f.

In both of the scenarios described above, requesting the labels of points in the region of disagreement resulted in a significant decrease in the region of disagreement. These would be considered *favorable* scenarios for disagreement-based active learning. However, we are not always so fortunate. For instance, consider again the scenario where a point is labeled "+" iff it is contained inside an unknown rectangle $[a_1, b_1] \times [a_2, b_2]$ in the plane, but this time suppose the data set ob-

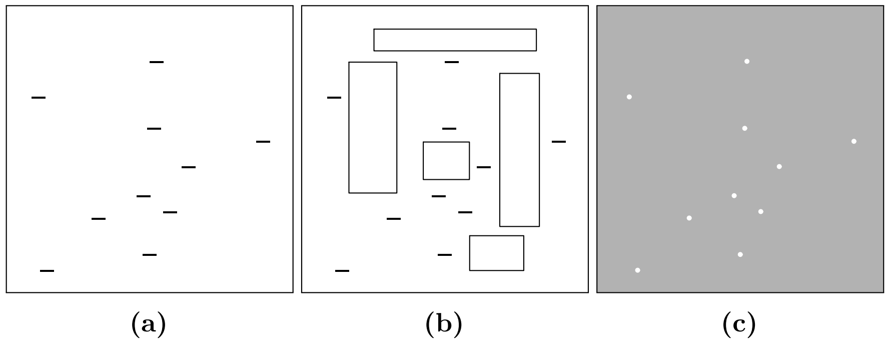

Figure 1.3: In the context of learning an axis-aligned rectangle, if all of the observed labels are "$-$", *every* point not in the data set is contained in the region of disagreement.

served so far is as depicted in Figure 1.3a. Note that all of the points in this data set are labeled "$-$". In this case, *every* rectangle that does not contain any of these data points would be consistent with their labels; a few such rectangles are depicted in Figure 1.3b. It should be clear that this is a very different kind of scenario from the prevous two. In particular, for every point (x_1, x_2) in the plane that is not among the few observed samples, the rectangle $[x_1, x_1] \times [x_2, x_2]$ containing *only this point* is consistent with all of the observed labels. Since this is true of *every* point not among the observed samples, the region of disagreement is the *entire space*, minus the few points in the data set; this is represented by the shaded region in Figure 1.3c. Thus, if we are given a new point that is not equal to one we have already observed the label of, the disagreement-based strategy will request its label. If, in response, we are told that the label is "$-$", then the region of disagreement is reduced by *only this single point*. In particular, if the probability distribution is non-atomic, then no matter how many samples labeled "$-$" we observe, the probability in the region of disagreement will always equal 1, and therefore *does not decrease*. Thus, if the unknown target rectangle has *zero* probability inside, then this situation will continue indefinitely (with probability 1), requesting every label and never reducing the probability in the region of disagreement.

The distinction raised by contrasting these two kinds of scenarios is fundamental to the active learning problem. In the chapters below, we will be highly interested in discussions of general conditions that distinguish between problems where the probability in the region of disagreement decreases (and approaches zero) and those where it does not. In the former case, we will be further interested in understanding the rates of decrease. With this understanding in hand, we are then able to describe the label complexities achieved by certain disagreement-based active learning algorithms *abstractly*. Various specific scenarios, such as those described above, can then be studied straightforwardly as special cases of the general analysis.

2

Basic Definitions and Notation

We begin by formalizing the active learning setting, defining the quantities that will be the focus of our discussion, and providing a few basic examples.

2.1 The Setting

We consider the following formal setting throughout this article. There is a set $\mathcal{X}$ called the *instance space*, equipped with a σ-algebra $\mathcal{B}_{\mathcal{X}}$; for convenience, let us suppose $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ is a standard Borel space (e.g., $\mathbb{R}^n$ under the usual Borel σ-algebra). Also let $\mathcal{Y} = \{-1, +1\}$, called the *label space*, and suppose $\mathcal{X} \times \mathcal{Y}$ is equipped with its product σ-algebra $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \otimes 2^{\mathcal{Y}}$. Fix a probability measure $\mathcal{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, called the *target distribution*, denote by $\mathcal{P}$ the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$, and $\forall x \in \mathcal{X}$, denote $\eta(x) = \mathbb{P}(Y = +1 | X = x)$, where $(X, Y) \sim \mathcal{P}_{XY}$. We refer to any measurable $h : \mathcal{X} \to \mathcal{Y}$ as a *classifier*. For any classifier h, define $\mathrm{er}(h) = \mathcal{P}_{XY}\left((x, y) : h(x) \neq y\right)$, called the *error rate*; in words, this is the probability that h makes a *mistake* in predicting the label Y by $h(X)$, for a random point $(X, Y) \sim \mathcal{P}_{XY}$. Throughout, let us make the usual simplifying assumption that all sets we evaluate the probabilities of, or functions we take expectations of, are indeed measurable; when this is not the case, one may typically turn to outer probabilities to maintain validity of the results, but we will not discuss these technical issues further below.

In this context, we are interested in learning from data: that is, producing a classifier h with small $\mathrm{er}(h)$, based on samples from $\mathcal{P}_{XY}$. Specifically, let $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of independent $\mathcal{P}_{XY}$-distributed random variables, called the *labeled data* sequence. For $m \in \mathbb{N}$, denote by $\mathcal{Z}_m = \{(X_i, Y_i)\}_{i=1}^{m}$ the first m data points. Also denote by $\mathcal{Z}_{\mathbf{X}} = \{X_i\}_{i=1}^{\infty}$ the *unlabeled data* sequence. Though in practice, the actual sequence of unlabeled data available would typically be large but finite, to focus our analysis on the number of *label requests* sufficient for learning, let us suppose we have access to the entire $\mathcal{Z}_{\mathbf{X}}$ sequence, representing an inexhaustible source of unlabeled data; the actual number of unlabeled data points needed by the algorithms below for their respective guarantees to hold can be extracted from their respective analyses.

In the *active learning* protocol, the learning algorithm is given a *budget* n, and provided direct access to $\mathcal{Z}_{\mathbf{X}}$. It may then select any index $i_1 \in \mathbb{N}$ and request to observe the label Y_{i_1}. Upon receiving the value of Y_{i_1}, it may then select another index i_2, request the label Y_{i_2}, and so on. After a number of these label requests not exceeding the budget n, the algorithm halts and returns a classifier $\hat{h}$. More formally, this protocol specifies a family of estimators that map $\mathcal{Z}$ to a classifier $\hat{h}$, such that for every $\mathcal{P}_{XY}$, $\hat{h}$ is conditionally independent of $\mathcal{Z}$ given $\mathcal{Z}_{\mathbf{X}}$ and $(i_1, Y_{i_1}), \ldots, (i_n, Y_{i_n})$, where each i_k is conditionally independent of $\mathcal{Z}$ given $\mathcal{Z}_{\mathbf{X}}$ and $(i_1, Y_{i_1}), \ldots, (i_{k-1}, Y_{i_{k-1}})$. In contrast, a *passive learning* algorithm is any (possibly randomized, independent from $\mathcal{Z}$) function $\mathcal{A}$ mapping a sequence $\mathcal{L} \in \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n$ of labeled data points to a classifier $\hat{h}$. We are then particularly interested in the behavior of $\mathcal{A}(\mathcal{Z}_n)$ as a function of n.

2.2 Basic Definitions

The primary focus in the study of active learning is the *label complexity*, defined formally as follows. A *label complexity* function Λ maps two values $\varepsilon, \delta \in [0,1]$ and a distribution $\mathcal{P}_{XY}$ to a value $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \in \mathbb{N} \cup \{\infty\}$.

Definition 2.1. For any active learning algorithm $\mathcal{A}$, we say $\mathcal{A}$ achieves label complexity Λ if, for every $\varepsilon \geq 0$ and $\delta \in [0,1]$, every distribution $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, and every integer $n \geq \Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$, if $\hat{h}$ is the classifier produced by running $\mathcal{A}$ with budget n, then with probability at least $1 - \delta$, $\mathrm{er}(\hat{h}) \leq \varepsilon$.

We will be particularly interested in the label complexity of achieving low error rate relative to the best error rate among a fixed set $\mathbb{C}$ of classifiers, known as the *hypothesis class*. In particular, denoting $\nu = \inf_{h \in \mathbb{C}} \mathrm{er}(h)$ (called the *noise rate*), we are typically interested in the value of $\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY})$ as a function of ε, δ, and $\mathcal{P}_{XY}$. For simplicity, we will suppose the infimum $\inf_{h \in \mathbb{C}} \mathrm{er}(h)$ is actually *achieved* by a classifier $f^\star \in \mathbb{C}$ (i.e., $\mathrm{er}(f^\star) = \nu$); otherwise, we could either let $f^\star \in \mathbb{C}$ be a classifier with $\mathrm{er}(f^\star)$ merely *close* to $\inf_{h \in \mathbb{C}} \mathrm{er}(h)$ [as done by Hanneke, 2011], or let $f^\star$ be in the *closure* of $\mathbb{C}$ with $\mathrm{er}(f^\star) = \nu$ [following Hanneke, 2012].

For comparison, we will also discuss the label complexity of certain passive learning algorithms $\mathcal{A}$. We can define this notion by considering a very simple type of active learning algorithm, which given budget n, simply requests the labels $Y_1, \ldots, Y_n$, and then returns the classifier produced by $\mathcal{A}(\mathcal{Z}_n)$. We then say $\mathcal{A}$ achieves a label complexity Λ under the same conditions specified by Definition 2.1, applied to this simple active learning algorithm.

Following the classic work of Vapnik and Chervonenkis [1971], for any $m \in \mathbb{N}$ and sequence $(x_1, \ldots, x_m) \in \mathcal{X}^m$, we say a set $\mathcal{H}$ of classifiers *shatters* $(x_1, \ldots, x_m)$ if, for every $(y_1, \ldots, y_m) \in \mathcal{Y}^m$, $\exists h \in \mathcal{H}$ s.t. $\forall i \in \{1, \ldots, m\}$, $h(x_i) = y_i$; in other words, $\mathcal{H}$ shatters $(x_1, \ldots, x_m)$ if all 2^m possible classifications of $(x_1, \ldots, x_m)$ can be realized by classifiers in $\mathcal{H}$. For convenience, define $\mathcal{X}^0 = \{()\}$ (where $()$ is the empty sequence), and say a set $\mathcal{H}$ shatters the empty sequence $()$ if and only if $\mathcal{H}$ is nonempty. The *Vapnik-Chervonenkis (VC) dimension* of a non-empty set $\mathcal{H}$, denoted $\mathrm{vc}(\mathcal{H})$, is defined as the largest integer m such that $\exists S \in \mathcal{X}^m$ shattered by $\mathcal{H}$, or as ∞ if no such value exists. We denote $d = \mathrm{vc}(\mathbb{C})$, and for simplicity, for the vast majority of the article, we will suppose $d < \infty$; in particular, many of the results below are stated in terms of d. We discuss other interesting scenarios, where d may be infinite, in Section 8.8.

For any set A, let $\mathbb{1}_A$ denote the indicator function for A: that is, $\mathbb{1}_A(x) = 1$ if $x \in A$, and $\mathbb{1}_A(x) = 0$ otherwise. We will also sometimes use the notation $\mathbb{1}[L]$, where L is a logical expression (e.g., "$f(x) \neq y$"), defining $\mathbb{1}[L] = 1$ if L is true, and $\mathbb{1}[L] = 0$ if L is false. Additionally, define the *signed indicator* function of A as $\mathbb{1}^{\pm}_A = 2\mathbb{1}_A - 1$. For a classifier h and a sequence of labeled data points $\mathcal{L} \in \bigcup_{m\in\mathbb{N}}(\mathcal{X} \times \mathcal{Y})^m$, define the *empirical error rate* of h with respect to $\mathcal{L}$ as $\mathrm{er}_{\mathcal{L}}(h) = \frac{1}{|\mathcal{L}|}\sum_{(x,y)\in\mathcal{L}} \mathbb{1}[h(x) \neq y]$, representing the fraction of points in $\mathcal{L}$ on which h makes mistakes. For completeness, also define $\mathrm{er}_{\emptyset}(h) = 0$. Also, when $\mathcal{L} = \mathcal{Z}_m$, the first m labeled data points, for any $m \in \mathbb{N} \cup \{0\}$, abbreviate $\mathrm{er}_m(h) = \mathrm{er}_{\mathcal{Z}_m}(h)$; also denote $V^\star_m = \{h \in \mathbb{C} : \forall i \leq m, h(X_i) = f^\star(X_i)\}$, called the *version space* induced by $\{X_1, \ldots, X_m\}$.

For any set of classifiers $\mathcal{H}$, and any $\varepsilon \in [0, 1]$, define the *ε-minimal set* as $\mathcal{H}(\varepsilon) = \{h \in \mathcal{H} : \mathrm{er}(h) - \inf_{g\in\mathcal{H}} \mathrm{er}(g) \leq \varepsilon\}$; also, for any classifier h, define the *ε-ball* centered at h as $\mathrm{B}_{\mathcal{H},\mathcal{P}}(h, \varepsilon) = \{g \in \mathcal{H} : \mathcal{P}(x : g(x) \neq h(x)) \leq \varepsilon\}$; when $\mathcal{H} = \mathbb{C}$, the hypothesis class, abbreviate $\mathrm{B}_{\mathcal{P}}(h, \varepsilon) = \mathrm{B}_{\mathbb{C},\mathcal{P}}(h, \varepsilon)$, and when $\mathcal{P}$ is clear from the context, abbreviate $\mathrm{B}_{\mathcal{H}}(h, \varepsilon) = \mathrm{B}_{\mathcal{H},\mathcal{P}}(h, \varepsilon)$, and $\mathrm{B}(h, \varepsilon) = \mathrm{B}_{\mathbb{C},\mathcal{P}}(h, \varepsilon)$. Additionally, define the *radius* of the set $\mathcal{H}$ as $\mathrm{radius}(\mathcal{H}) = \sup_{h\in\mathcal{H}} \mathcal{P}(x : h(x) \neq f^\star(x))$, which is the smallest ε for which $\mathcal{H} = \mathrm{B}_{\mathcal{H}}(f^\star, \varepsilon)$. Finally, define the *region of disagreement* of $\mathcal{H}$ as

$$\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\},$$

the set of points for which there is some disagreement among classifiers in $\mathcal{H}$ regarding their predicted label.

Below, we will study a certain family of active learning algorithms, based on a general strategy known as *disagreement-based* active learning [Cohn, Atlas, and Ladner, 1994, Balcan, Beygelzimer, and Langford, 2006]. This strategy involves maintaining a set V of candidate classifiers (one of which will be returned in the end), processing the unlabeled samples in sequence, and requesting the labels Y_i of samples X_i in $\mathrm{DIS}(V)$. This ensures that we request the label of any sample for which there is some uncertainty about the classification the returned classifier will assign to it. The set V is then periodically updated by removing classifiers with relatively poor performance on the queried samples. We will discuss this strategy in more detail in Chapter 5, but even from this rough description, it should be clear that analysis of its label complexity will necessarily involve characterizing properties of the regions $\mathrm{DIS}(V)$, for the sets V obtained in the course of the execution. In particular, since this strategy only requests the labels of samples in $\mathrm{DIS}(V)$, it will be important to characterize the probability that a random sample X_i is in $\mathrm{DIS}(V)$: that is, $\mathcal{P}(\mathrm{DIS}(V))$.

As we will see below, it is often straightforward to express a concise bound on $\mathrm{radius}(V)$, for the sets V obtained in these algorithms. For this reason, in the interest of obtaining concise bounds on the label complexity, it will often be convenient to bound $\mathcal{P}(\mathrm{DIS}(V))$ by a homogeneous linear function of a bound on $\mathrm{radius}(V)$. In the context of active learning, the coefficient in this linear function is typically referred to as the *disagreement coefficient* [following Hanneke, 2007b, 2009b]. A nearly-identical quantity has also appeared in the literature on ratio-type empirical processes [Alexander, 1987, Giné and Koltchinskii, 2006], there typically referred to as *Alexander's capacity function*. In both of these contexts, it is essentially used to describe the rate of collapse of $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, \varepsilon)))$ as $\varepsilon \to 0$. It is formally defined as follows.

Definition 2.2. For any $r_0 \geq 0$ and classifier h, define the *disagreement coefficient* of h with respect to $\mathbb{C}$ under $\mathcal{P}$ as

$$\theta_h(r_0) = \sup_{r > r_0} \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(h, r\right)\right)\right)}{r} \vee 1.$$

When $h = f^\star$, abbreviate this as $\theta(r_0) = \theta_{f^\star}(r_0)$, called the disagreement coefficient of the class $\mathbb{C}$ with respect to $\mathcal{P}_{XY}$.

Recalling the motivating discussion above, note that, for any $V \subseteq \mathbb{C}$ and $r \geq \max\{\mathrm{radius}(V), r_0\}$, we have $\mathcal{P}(\mathrm{DIS}(V)) \leq \theta(r_0) r$, so that the disagreement coefficient can indeed be used to relate $\mathcal{P}(\mathrm{DIS}(V))$ to radius(V). In general, the value $\theta_h(r_0)$ can always be upper-bounded by $\theta_h(0)$, or even $\sup_h \theta_h(0)$. However, we will see below that in many scenarios, $\theta_h(r_0)$ exhibits more-interesting behaviors if we maintain the dependence on h and r_0 (see Section 2.4). In particular, note that since probabilities are never greater than 1, for any $r_0 > 0$ we always have $1 \leq \theta_h(r_0) \leq 1/r_0$.

We go through several simple examples of calculating $\theta_h(r_0)$ in detail in Section 2.4 below, and several more-sophisticated examples in Chapter 7. As a simple illustration for now, consider the case of *threshold* classifiers (Example 1 below), in which $\mathcal{X} = [0,1]$ and $\mathbb{C} = \{\mathbb{1}^{\pm}_{[z,1]} : z \in (0,1)\}$, and suppose $\mathcal{P}$ is the uniform distribution over $[0,1]$. In this case, for $h = \mathbb{1}^{\pm}_{[z,1]} \in \mathbb{C}$ and $r > r_0$, $\mathrm{B}(h,r) = \{\mathbb{1}^{\pm}_{[z',1]} : z' \in [z-r, z+r] \cap (0,1)\}$, $\mathrm{DIS}(\mathrm{B}(h,r)) = [z-r, z+r) \cap (0,1)$, and thus $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r))) \leq 2r$, so that $\theta_h(r_0) = \sup_{r>r_0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))/r \leq 2$. Furthermore, for all sufficiently small r, $\mathrm{DIS}(\mathrm{B}(h,r)) = [z-r, z+r)$, in which case $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r))) = 2r$; therefore, $\theta_h(0) = 2$.

Before proceeding, we should clarify the use of certain asymptotic notation appearing below. Specifically, we make use of the standard "O" notation, for functions of $\varepsilon \in (0,1)$, as well as functions of $n \in \mathbb{N}$. Generally, when considering a function of some variable $\varepsilon \in (0,1)$, the asymptotics are considered as $\varepsilon \to 0$; when considering a function of a variable $n \in \mathbb{N}$, the asymptotic behavior is exhibited as $n \to \infty$. Additionally, for any statement of the form "$x \to 0$", we always mean that the limit is taken *from above*: that is, $x \downarrow 0$. For example, for functions $u : (0,1) \to [0,\infty]$ and $v : (0,1) \to [0,\infty]$, the statement that $\limsup\limits_{\varepsilon \to 0} \frac{u(\varepsilon)}{v(\varepsilon)} < \infty$ can be equivalently expressed as either $u(\varepsilon) = O(v(\varepsilon))$ or $v(\varepsilon) = \Omega(u(\varepsilon))$; we say $u(\varepsilon) = \Theta(v(\varepsilon))$ if $u(\varepsilon) = O(v(\varepsilon))$ and $u(\varepsilon) = \Omega(v(\varepsilon))$. Likewise, the statement that $\lim\limits_{\varepsilon \to 0} \frac{u(\varepsilon)}{v(\varepsilon)} = 0$ can be equivalently expressed as either $u(\varepsilon) = o(v(\varepsilon))$ or $v(\varepsilon) = \omega(u(\varepsilon))$. We also use the standard notation for asymptotics involving sets; for instance, for a monotone collection of sets $\{A_r\}_{r \in [0,\infty)}$, the set $\lim_{r \to 0} A_r$ is defined by the property that $\mathbb{1}_{\lim\limits_{r \to 0} A_r} = \lim\limits_{r \to 0} \mathbb{1}_{A_r}$. We will also make use of a non-asymptotic notation for inequalities when, for simplicity, we refrain from expressing explicit numerical constant factors. Specifically, the relation $u(\varepsilon, \delta) \lesssim v(\varepsilon, \delta)$ denotes the statement that there exists a universal numerical constant $c \in (0, \infty)$ (i.e., having no dependence on the particular $\mathbb{C}$, $\mathcal{P}_{XY}$, or other such problem-specific variables) such that $u(\varepsilon, \delta) \leq cv(\varepsilon, \delta)$ for all $\varepsilon, \delta \in (0, 1)$. Likewise, we will sometimes write $u(\varepsilon, \delta) \gtrsim v(\varepsilon, \delta)$ to denote that $u(\varepsilon, \delta) \geq cv(\varepsilon, \delta)$ for all $\varepsilon, \delta \in (0, 1)$, where $c \in (0, \infty)$ is some implicit universal constant.

One additional notational convention we will use throughout is that, for $x \geq 0$, we denote $\mathrm{Log}(x) = \max\{\ln(x), 1\}$.

2.3 Noise Models

We will formulate the results below in terms of a few commonly-studied noise conditions. Our most general results below will hold for *any* distribution $\mathcal{P}_{XY}$, and will typically be stated in terms of the *noise rate* $\nu = \mathrm{er}(f^\star)$. However, as we will see, we can exhibit more interesting behaviors in label complexities under a more detailed description of $\mathcal{P}_{XY}$, stated in the following condition.

Condition 2.3. For some $a \in [1, \infty)$ and $\alpha \in [0, 1]$, for every $h \in \mathbb{C}$,

$$\mathcal{P}\left(x : h(x) \neq f^\star(x)\right) \leq a\left(\mathrm{er}(h) - \mathrm{er}(f^\star)\right)^\alpha.$$

$\mathcal{P}_{XY}$ will always satisfy Condition 2.3 with $\alpha = 0$ (supposing we interpret $0^0 = 1$ in this context); however, our primary interest in the present article will be scenarios in which this condition is satisfied with $\alpha > 0$. This type of condition was first introduced by Mammen and Tsybakov [1999] and Tsybakov [2004], and is referred to in the literature variously as *Tsybakov noise*, *Mammen-Tsybakov noise*, a *margin condition*, or as a *low noise* condition. There is now an extensive literature on the achievable label complexities (both for passive and active) under this condition [e.g., Massart and Nédélec, 2006, Koltchinskii, 2006, Bartlett, Jordan, and McAuliffe, 2006, Castro and Nowak, 2008, Wang, 2011, Koltchinskii, 2010, Hanneke, 2011, 2012]. Mammen and Tsybakov [1999] show that, to satisfy Condition 2.3 with some $\alpha \in (0, 1)$, it suffices that $f^\star$ is the Bayes optimal classifier (i.e., the global minimizer of $\mathrm{er}(h)$ over all possible classifiers

h), and $\forall t > 0$,

$$\mathcal{P}\left(x : |\eta(x) - 1/2| \leq t\right) \leq a' t^{\alpha/(1-\alpha)}, \tag{2.1}$$

where $a' = (1-\alpha)(2\alpha)^{\alpha/(1-\alpha)} a^{1/(1-\alpha)}$ [see also Tsybakov, 2004]; this can be interpreted as saying that the probability of X with high-entropy conditional $Y|X$ is small. For this reason, (2.1) is often referred to as a *low noise* condition. Furthermore, to satisfy Condition 2.3 with $\alpha = 1$ and a given value of $a \in [1, \infty)$, it suffices to have $f^\star$ equal the Bayes optimal classifier, and

$$\mathcal{P}\left(x : |\eta(x) - 1/2| < 1/(2a)\right) = 0, \tag{2.2}$$

which is referred to as a *bounded noise* condition, or sometimes as *Massart noise* [Massart and Nédélec, 2006, Giné and Koltchinskii, 2006].

This condition can be realized in a variety of ways, yielding complementary interpretations. For instance, for certain hypothesis classes with a kind of geometric interpretation (e.g., linear separators), Condition 2.3 can often be interpreted as a way to relate $\eta(x)$, the density of $\mathcal{P}$ at x, and the distance of x to the decision boundary of $f^\star$. That is, suppose $\eta(x)$ approaches $1/2$ as x approaches the $f^\star$ decision boundary. If $\mathcal{P}$ has high density around that decision boundary, then the value of α in Condition 2.3 can often be interpreted as indicating the *rate* at which $\eta(x)$ approaches $1/2$ as a function of the distance of x to the decision boundary [see Castro and Nowak, 2008]. On the other hand, if we fix the form of $\eta(x)$ as a function of the distance from x to the $f^\star$ decision boundary, then the value of α in Condition 2.3 may often be interpreted as a kind of margin condition on $\mathcal{P}$, specifying the rate at which the density of $\mathcal{P}$ at x vanishes as x approaches the decision boundary of $f^\star$ [see Cavallanti, Cesa-Bianchi, and Gentile, 2011, Dekel, Gentile, and Sridharan, 2012].

There is a special case of Condition 2.3 that is of particular interest, primarily due to the simplicity and elegance it admits in the development of algorithms and their analysis: namely, the *realizable case*. Specifically, we say $\mathcal{P}_{XY}$ is in the *realizable case* if $\mathrm{er}(f^\star) = 0$. That is, $f^\star$ is essentially flawless. In particular, without loss, in the realizable case we can suppose $Y_i = f^\star(X_i)$ for every $i \in \mathbb{N}$, since this is the case with probability 1. Due to this special property, in the realizable case we typically refer to $f^\star$ as the *target function*, indicating that it represents the concept to be learned. The realizable case is a classic scenario studied extensively in the computational learning theory literature, most often in the context of the so-called *PAC model* of Valiant [1984]. For our purposes, it will serve as a kind of staging ground, a much simpler setting in which to develop active learning methods and the analyses there-of, without the need to worry about certain issues that come up when there are noisy labels. As we will see, for the techniques we focus on here, much of the intuition we develop for the realizable case carries over to the noisy case, and in particular, there is an (almost mechanical) technique for making such methods robust to noise via minor modifications to the algorithms and the analysis there-of.

2.4 Basic Examples

Before continuing with the general analysis below, we first introduce a few basic examples: namely, threshold classifiers, interval classifiers, and linear separators. We repeatedly refer to these examples throughout the article. The first two of these should essentially be considered *toy* problems, studied primarily to exhibit certain issues that arise in the analysis of active learning in their barest forms. The last of these examples presently represents the most-studied (non-toy) active learning problem, and we will discuss it in detail in later sections. Several additional examples are provided in later sections as well, to illustrate various issues and behaviors discussed there.

Example 1 As mentioned above, in the problem of learning *threshold* classifiers, we take $\mathcal{X} = [0, 1]$ and $\mathbb{C} = \{\mathbb{1}^{\pm}_{[z,1]} : z \in (0, 1)\}$.

The problem of learning a threshold classifier is perhaps the clearest example of how active learning can provide significant benefits over passive learning in terms of label complexity. One simple passive learning algorithm for the realizable case would, when given as input $\mathcal{Z}_n$, simply return the classifier $\hat{h} = \mathbb{1}^{\pm}_{[\hat{z},1]}$, where $\hat{z}$ is the midpoint between $\max\{X_i : Y_i = -1, 1 \leq i \leq n\} \cup \{0\}$ and $\min\{X_i : Y_i = +1, 1 \leq i \leq n\} \cup \{1\}$. Supposing $\mathcal{P}$ is the uniform distribution on $\mathcal{X}$, and $f^\star = \mathbb{1}^{\pm}_{[z^\star,1]}$, where $\varepsilon < z^\star < 1-\varepsilon$, to guarantee $\mathrm{er}(\hat{h}) \leq \varepsilon$, it suffices to have some $X_i \in [z^\star - \varepsilon, z^\star)$ and another $X_i \in [z^\star, z^\star + \varepsilon]$. Each of these regions has probability ε, so the probability this happens is at least $1 - 2(1-\varepsilon)^n$ (by a union bound); since $1 - \varepsilon \leq e^{-\varepsilon}$, this is at least $1 - 2e^{-\varepsilon n}$. For this to be greater than $1 - \delta$, it suffices to take $n \geq \frac{1}{\varepsilon} \ln \frac{2}{\delta}$. Similar reasoning applies to $z^\star \in [0, \varepsilon) \cup (1-\varepsilon, 1]$ (in which case only one of these regions needs an X_i point in it), so that for $\mathcal{P}_{XY}$ in the realizable case, this passive learning algorithm achieves a label complexity $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) = \left\lceil \frac{1}{\varepsilon} \ln \frac{2}{\delta} \right\rceil$. Furthermore, it is easy to see that *any* label complexity Λ achieved by a passive learning algorithm must have some $\mathcal{P}_{XY}$ in the realizable case with $\mathcal{P}$ uniform over $\mathcal{X}$ and $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \gtrsim \frac{1}{\varepsilon} \ln \frac{1}{\delta}$ (e.g., this must be true either for $f^\star = \mathbb{1}^{\pm}_{[\varepsilon/2,1]}$ or $f^\star = \mathbb{1}^{\pm}_{[3\varepsilon,1]}$, since we would need a point in $[\varepsilon/2, 3\varepsilon)$ to distinguish between these cases), so that this is a fairly reasonable analysis of the capabilities of passive learning for this problem.

On the other hand, consider the following simple active learning algorithm, when given budget n. Let $m = 2^{n-1}$ and let $\{j_k\}_{k=1}^m$ be a sequence of distinct indices in $\{1, \ldots, m\}$ such that $X_{j_1} \leq X_{j_2} \leq \cdots \leq X_{j_m}$ (i.e., the sorted order of $\{X_1, \ldots, X_m\}$). Also initialize $\ell = 0$ and $u = m+1$. Then repeat the following steps until $\ell = u - 1$ is satisfied: let $k = \lfloor (\ell + u)/2 \rfloor$, request the label Y_{j_k} of the point X_{j_k}; if $Y_{j_k} = +1$, let $u = k$, and otherwise let $\ell = k$. Once we have $\ell = u - 1$, we return $\hat{h} = \mathbb{1}^{\pm}_{[\hat{z},1]}$, where $\hat{z} = (X_{j_\ell} + X_{j_u})/2$ if $\ell > 0$ and $u < m+1$, or $\hat{z} = X_{j_u}/2$ if $\ell = 0$, or $\hat{z} = (X_{j_\ell} + 1)/2$ if $u = m+1$.

First note that, since k is the median between ℓ and u, and either ℓ or u is set to k after each label request, the total number of label requests is at most $\log_2(m) + 1 = n$, so that this algorithm indeed stays within the indicated budget. Second, note that the algorithm maintains the invariant that either $\ell = 0$ or X_{j_ℓ} is the largest point among $\{X_1, \ldots, X_m\}$ for which Y_{j_ℓ} has been requested and observed to be -1, and also that either $u = m+1$ or X_{j_u} is the smallest point among $\{X_1, \ldots, X_m\}$ for which Y_{j_u} has been requested and observed to be $+1$. In particular, this means that in the realizable case, at the end every $k \in \{u, \ldots, m\}$ has $Y_{j_k} = +1$, and every $k \in \{1, \ldots, \ell\}$ has $Y_{j_k} = -1$. Since $\ell = u - 1$ in the end, we see that $\hat{h}$ is precisely the same as the classifier that would be produced by the above passive learning algorithm when given $\mathcal{Z}_m$ as input. This is remarkable, since m is exponentially larger than n. In particular, this immediately implies that this active learning algorithm achieves a label complexity Λ that, for $\mathcal{P}_{XY}$ in the realizable case, satisfies $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \leq 1 + \left\lceil \log_2 \left(\frac{1}{\varepsilon} \ln \frac{2}{\delta} \right) \right\rceil$, which is an exponential improvement over passive learning.

As we will see, this logarithmic dependence on $1/\varepsilon$ is typically the best we can hope for in active learning, and it will continue to be available under Condition 2.3 with $\alpha = 1$, though not for $\alpha < 1$. The $\log\log(1/\delta)$ dependence on $1/\delta$ in this example is also somewhat interesting; in fact, via algorithms that make use of larger quantities of unlabeled data, the dependence on δ can be entirely eliminated from the label complexity bound. Just how general this phenomenon is in the realizable case has not yet been fully explored in the literature. However, unlike the logarithmic dependence on $1/\varepsilon$, we will see that this improved dependence on $1/\delta$ is typically *not* available under interesting noise models.

For the problem of learning threshold classifiers, it is easy to see that $\mathrm{vc}(\mathbb{C}) = 1$, since any $x \in (0, 1)$ has $\{x\}$ shatterable, while for any $x' \leq x$, no $h \in \mathbb{C}$ realizes the labeling $\{(x', +1), (x, -1)\}$. As discussed above, it is also easy to bound $\theta_h(\varepsilon)$ for threshold classifiers; in particular, for $\mathcal{P}$ the uniform distribution on $\mathcal{X}$, we found that $\theta_h(\varepsilon) \leq 2$ for any $h \in \mathbb{C}$. In fact, a careful examination reveals that, for $h = \mathbb{1}^{\pm}_{[z,1]} \in \mathbb{C}$ and $\varepsilon \in (0, 1)$, we have precisely $\theta_h(\varepsilon) = \min\{1, z/\varepsilon\} + \min\{1, (1 - z)/\varepsilon\} \in [1, 2]$, and in particular $\theta_h(0) = 2$. Furthermore, it is straightforward to extend the upper-bounding argument to any distribution $\mathcal{P}$ over $\mathcal{X}$, where more generally, we have $\mathrm{B}(h, r) = \{\mathbb{1}^{\pm}_{[z',1]} \in \mathbb{C} : \mathcal{P}([\min\{z, z'\}, \max\{z, z'\}]) \leq r\}$. In this case, we again have $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r))) \leq 2r$. Thus, threshold classifiers have $1 \leq \theta_h(\varepsilon) \leq 2$ for all distributions.

Example 2 In the problem of learning *interval* classifiers, $\mathcal{X} = [0,1]$ and $\mathbb{C} = \{\mathbb{1}^{\pm}_{[a,b]} : 0 < a \leq b < 1\}$.

While the analysis of learning interval classifiers by passive learning remains almost identical to that for threshold classifiers, the issues arising from the analysis of active learning are far more subtle for this space. Specifically, we can design a passive learning algorithm which, given $\mathcal{Z}_n$ as input, returns $\hat{h} = \mathbb{1}^{\pm}_{[\hat{a},\hat{b}]}$, where $\hat{a}$ is the smallest X_i for which $i \leq n$ and $Y_i = +1$, and $\hat{b}$ is the largest X_i with $i \leq n$ and $Y_i = +1$, or else $\hat{h} = -1$ if no $Y_i = +1$ with $i \leq n$; this is known as the *Closure* algorithm [Helmbold, Sloan, and Warmuth, 1990]. For $\mathcal{P}$ uniform over $\mathcal{X}$, $\mathcal{P}_{XY}$ in the realizable case, and $f^\star = \mathbb{1}^{\pm}_{[a,b]}$, where $b - a > \varepsilon$, to guarantee $\mathrm{er}(\hat{h}) \leq \varepsilon$, it suffices to have some $X_i \in [a, a+\varepsilon/2]$ and some $X_i \in [b - \varepsilon/2, b]$, where $i \leq n$ in both cases. Each of these regions has probability $\varepsilon/2$, so that the probability this happens is at least $1 - 2(1-\varepsilon/2)^n \geq 1 - 2e^{-n\varepsilon/2}$; thus, any $n \geq \frac{2}{\varepsilon}\ln\frac{2}{\delta}$ suffices to guarantee this happens with probability at least $1 - \delta$. If $b - a \leq \varepsilon$, then it is easy to see that this algorithm *always* has $\mathrm{er}(\hat{h}) \leq \varepsilon$. Thus, in general, this algorithm achieves a label complexity Λ that, for $\mathcal{P}_{XY}$ in the realizable case with $\mathcal{P}$ uniform, satisfies $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \leq \left\lceil \frac{2}{\varepsilon}\ln\frac{2}{\delta} \right\rceil$. Again, one can show that this is a fairly tight characterization of the capabilities of passive learning algorithms for this problem in general, as there are known lower bounds of a similar form.

Now let us examine the label complexities achievable by active learning, supposing again that $\mathcal{P}_{XY}$ is in the realizable case, and that $\mathcal{P}$ is uniform over $\mathcal{X}$. For simplicity, in this example (only) let us suppose the active learning algorithm can request the label $f^\star(x)$ of *any* $x \in [0,1]$; this is a mild assumption for this problem, since $\mathcal{P}$ being uniform implies that we can (with probability 1) find points X_i in $\mathcal{Z}_{\mathbf{X}}$ arbitrarily close to any such x. We propose an active learning algorithm with two stages. In the first stage, the algorithm requests the labels of points x on a sequence of increasingly-refined grids: specifically, it requests the labels of the points $1/2, 1/4, 3/4, 1/8, 3/8, 5/8, 7/8, 1/16, 3/16, \ldots$, sequentially. This continues until either the number of label requests meets the budget n, or the response to some label request indicates the requested label is $+1$. In the former case, the algorithm simply returns $\hat{h} = -1$. In the latter case, the algorithm enters a second stage; letting x_+ denote the x corresponding to this first returned positive data point, the algorithm initializes $\ell_1 = 0$, $u_1 = x_+$, $\ell_2 = x_+$, and $u_2 = 1$. Supposing it has used k label requests up to this point, it then repeats the following $\lfloor (n-k)/2 \rfloor$ times: request the label of $x = (\ell_1 + u_1)/2$; if the response is $+1$, let $u_1 = x$, and otherwise let $\ell_1 = x$. After this, it similarly repeats the following $\lceil (n-k)/2 \rceil$ times: request the label of $x = (\ell_2 + u_2)/2$; if the response is $+1$, let $\ell_1 = x$, and otherwise let $u_1 = x$. In the end, it returns the classifier $\hat{h} = \mathbb{1}^{\pm}_{[u_1,\ell_2]}$.

As was the case for thresholds, the second stage of this algorithm maintains the invariant that ℓ_1 and u_1 are the closest negative and positive points, respectively, that are less than x_+, while ℓ_2 and u_2 are the closest positive and negative points, respectively, that are greater than x_+. The distance between ℓ_1 and u_1 is halved after each label request in the first repeated step, while the same is true of ℓ_2 and u_2 in the second repeated step. Note that if $f^\star = \mathbb{1}^{\pm}_{[a,b]}$ and $b - a \le \varepsilon$, then this algorithm *always* has $\operatorname{er}(\hat{h}) \le \varepsilon$. Otherwise, if $b - a > \varepsilon$, then the number of label requests before encountering the first positive label (or halting) will not exceed $\left\lceil \frac{2}{b-a} \right\rceil$. At that point, if the remaining budget is at least $2\lceil \log_2(2/\varepsilon) \rceil$, then the second phase will produce ℓ_1, u_1, ℓ_2, u_2 with $u_1 - \ell_1 \le \varepsilon/2$ and $u_2 - \ell_2 \le \varepsilon/2$; since $\ell_1 \le a \le u_1$ and $\ell_2 \le b \le u_2$, this clearly implies $\operatorname{er}(\hat{h}) \le \varepsilon$. Thus, this algorithm achieves label complexity Λ such that, for $\mathcal{P}_{XY}$ in the realizable case with $\mathcal{P}$ uniform, letting $w = \mathcal{P}(x : f^\star(x) = +1)$, $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \le \left\lceil \frac{2}{\max\{w,\varepsilon\}} \right\rceil + 2\lceil \log_2(2/\varepsilon) \rceil$.

There are several ways to interpret this result. On a first look, we might conclude that those $\mathcal{P}_{XY}$ with $\mathcal{P}(x : f^\star(x) = +1)$ much larger than ε have a large improvement over the passive label complexity, while those with $\mathcal{P}(x : f^\star(x) = +1)$ close to ε have a label complexity quite similar to that of the passive learning algorithm. This is indeed a valid observation. However, there is also an alternative perspective on this result, arising from examining the asymptotic behavior as $\varepsilon \to 0$. Specifically, for any $\mathcal{P}_{XY}$ with $w = \mathcal{P}(x : f^\star(x) = +1) > 0$, considering the distribution $\mathcal{P}_{XY}$ as *fixed*, the asymptotic dependence on ε is $O(\log(1/\varepsilon))$. Since, in the case $w = 0$, the algorithm *always* produces $\hat{h}$ with $\mathrm{er}(\hat{h}) \leq \varepsilon$, we see that this algorithm achieves a label complexity that, for every $\mathcal{P}_{XY}$ in the realizable case with $\mathcal{P}$ uniform, satisfies $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) = O(\log(1/\varepsilon))$. It is also straightforward to extend this result to nonuniform $\mathcal{P}$ distributions. As we discuss in Chapter 3, passive learning algorithms typically cannot achieve label complexities with sublinear dependence on $1/\varepsilon$, even in this asymptotic sense as $\varepsilon \to 0$ with $\mathcal{P}_{XY}$ fixed, so that this $O(\log(1/\varepsilon))$ dependence on ε may be regarded as a strong improvement over passive learning. Thus, in this example, though in some sense it is quite easy to produce a negative result about the label complexity of active learning (in this case for small intervals), the negative result only arises for values of ε that are large relative to some quantity related to $\mathcal{P}_{XY}$ (in this case, w), and vanish as $\varepsilon \to 0$. This phenomenon will come up repeatedly in the analysis below. In particular, this highlights the need to express label complexity bounds in active learning in terms of $\mathcal{P}_{XY}$-dependent quantities, so that both interpretations are implicit in each of the results.

As was the case for thresholds, we can easily calculate $\mathrm{vc}(\mathbb{C})$ and $\theta_h(\varepsilon)$ for intervals. Specifically, any $x < x'$ can be shattered (by taking $\mathbb{1}^{\pm}_{[a,b]}$ with $a \in \{x, x+\varepsilon\}$ and $b \in \{x'-\varepsilon, x'\}$, for $2\varepsilon < x' - x$), while for any $x \leq x' \leq x''$, $\mathbb{C}$ cannot realize the labeling $\{(x, +1), (x', -1), (x'', +1)\}$; thus, $\mathrm{vc}(\mathbb{C}) = 2$. We can also calculate $\theta_h(\varepsilon)$ for $h \in \mathbb{C}$, and $\mathcal{P}$ the uniform distribution. For any $h = \mathbb{1}^{\pm}_{[a,b]} \in \mathbb{C}$, and $r < b - a$, $\mathrm{B}(h, r) = \{\mathbb{1}^{\pm}_{[a',b']} \in \mathbb{C} : |a' - a| + |b' - b| \leq r\}$, so that $\mathrm{DIS}(\mathrm{B}(h, r)) = ([a-r, a+r] \cup (b-r, b+r]) \cap (0, 1)$, and $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r))) \leq 4r$. For $r > b - a$, every $a' \in (0, 1)$ has $\mathbb{1}^{\pm}_{[a',a']} \in \mathrm{B}(h, r)$, so that $\mathrm{DIS}(\mathrm{B}(h, r)) = (0, 1)$, and $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r))) = 1$. Since $\theta_h(\varepsilon)$ is always at most $1/\varepsilon$, we have $\theta_h(\varepsilon) \leq \max\left\{\frac{1}{\max\{b-a,\varepsilon\}}, 4\right\}$. Furthermore, as in Example 1, this inequality becomes an equality for all sufficiently small ε. It is straightforward to generalize this to arbitrary distributions $\mathcal{P}$, in which case $\theta_h(\varepsilon) \leq \max\left\{\frac{1}{\max\{\mathcal{P}(x:h(x)=+1),\varepsilon\}}, 4\right\}$ for any $h \in \mathbb{C}$. Note that, unlike thresholds, the disagreement coefficient $\theta_h(\varepsilon)$ with respect to interval classifiers has a nontrivial dependence on the classifier h, so that it exhibits an interesting range of behaviors.

Example 3 Another important example for active learning, and machine learning in general, is the class of *k-dimensional linear separators.* Formally, in the problem of learning k-dimensional linear separators, where $k \in \mathbb{N}$, we have $\mathcal{X} = \mathbb{R}^k$, and $\mathbb{C} = \{\mathbb{1}^{\pm}_{\{x: w^T x + b \geq 0\}} : b \in \mathbb{R}, w \in \mathbb{R}^k, \|w\| > 0\}$. In other words, for each classifier $h \in \mathbb{C}$, there is an associated *weight vector* $w = (w_1, \ldots, w_k) \in \mathbb{R}^k \setminus \{0^k\}$ and *bias* $b \in \mathbb{R}$, and any $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$ has $h(x) = +1$ if and only if $b + \sum_{i=1}^k w_i x_i \geq 0$. The set $\{x : w^T x + b = 0\}$ is referred to as the *separating hyperplane.* For technical reasons, in this article, we do not include weight vectors w with $\|w\| = 0$, since this would introduce discontinuities and asymmetries that would complicate the discussion. It is not difficult to show that $\mathrm{vc}(\mathbb{C}) = k + 1$ [see e.g., Cover, 1965, Vapnik and Chervonenkis, 1971, Devroye, Györfi, and Lugosi, 1996]. We will sometimes refer to the class of *homogeneous* linear separators, which is simply the subclass having $b = 0$; equivalently, homogeneous linear separators are those for which the separating hyperplane passes through the origin.

The class of linear separators is presently the most commonly-used hypothesis class in machine learning applications, and moreover, it is by far the most commonly-used hypothesis class in the literature on applications of active learning. As such, it has also received considerable attention in the theoretical active learning literature, and we will discuss this class in substantial detail in this article. In particular, we will study the disagreement coefficients $\theta_h(\varepsilon)$ with respect to this class under various conditions on h and $\mathcal{P}$ in Chapter 7.

3

A Brief Review of Passive Learning

Before proceeding with the general discussion of active learning, we first review the known results on the label complexity of passive learning. These results will serve as reference points for comparison, so that we can judge the relative improvements in the label complexity of active learning, compared to passive learning, in later sections. They will also be directly useful for us, in the context of the design and analysis of the active learning algorithms presented in later sections, since certain steps in the algorithms will make use of the same concentration inequalities that play a fundamental role in the analysis of passive learning methods. However, since the main focus of this article is active learning, I do not provide proofs of the results here. The interested reader is referred to the respective referenced literature for each result below.

3.1 General Concentration Inequalities

The most common approach to the analysis of the label complexity of passive learning is by concentration inequalities, rooted in the classic works of Vapnik and Chervonenkis [1971] and Vapnik [1982], bounding the largest difference between excess error rate and excess empirical error rate among classifiers in $\mathbb{C}$. Specifically, the following lemma presents one such bound, which we refer to repeatedly throughout this article. This particular result follows from the work of Giné and Koltchinskii [2006] (slightly refining an earlier theorem of Massart and Nédélec, 2006); see the work of Hanneke and Yang [2012] for an explicit derivation, from which this lemma easily follows.

Lemma 3.1. There is a universal constant $c \in (1, \infty)$ such that, for any $\gamma \in (0, 1)$, and any $m \in \mathbb{N}$, letting $\varepsilon_m = \left(\frac{ad}{m}\right)^{\frac{1}{2-\alpha}}$, and

$$U(m, \gamma) = c \min \begin{cases} \left(\frac{a(d\mathrm{Log}(\theta(a\varepsilon_m^\alpha)) + \mathrm{Log}(1/\gamma))}{m}\right)^{\frac{1}{2-\alpha}} \\ \frac{d\mathrm{Log}(\theta(d/m)) + \mathrm{Log}(1/\gamma)}{m} + \sqrt{\frac{\nu(d\mathrm{Log}(\theta(\nu)) + \mathrm{Log}(1/\gamma))}{m}} \end{cases},$$

with probability at least $1 - \gamma$, $\forall h \in \mathbb{C}$, the following inequalities hold:

$$\begin{aligned} \mathrm{er}(h) - \mathrm{er}(f^\star) &\leq \max\left\{2\left(\mathrm{er}_m(h) - \mathrm{er}_m(f^\star)\right), U(m, \gamma)\right\}, \\ \mathrm{er}_m(h) - \min_{g \in \mathbb{C}} \mathrm{er}_m(g) &\leq \max\left\{2\left(\mathrm{er}(h) - \mathrm{er}(f^\star)\right), U(m, \gamma)\right\}. \end{aligned}$$

Readers acquainted with the learning theory literature may be more familiar with results of this type having "m/d" in place of $\theta(d/m)$ and $\theta(\nu)$ above [e.g., Vapnik, 1982, 1998], or with "$a^{-2/\alpha}m/d$" in place of $\theta(a\varepsilon_m^\alpha)$ [Massart and Nédélec, 2006]. These well-known results follow easily from this one, by the basic fact that $\theta(\varepsilon) \leq 1/\varepsilon$, and by noting that the term depending on ν is dominated by the term to its left if $\nu < d/m$. One could potentially use those more-familiar bounds in place of $U(m, \gamma)$ in each of the instances below; the expense of doing so would be an increase by at most a logarithmic factor in the corresponding label complexity results.

We will often be interested in determining a sufficient number m of samples to guarantee $U(m, \gamma)$ is below a given size. Using basic properties of the disagreement coefficient (namely, Corollary 7.2) and a bit of algebra, it is straightforward to show that, for some universal constant $c' \in [1, \infty)$, for any $m \in \mathbb{N}$ and $\varepsilon, \gamma \in (0, 1)$,

$$m \geq c' a \varepsilon^{\alpha-2} \left(d\mathrm{Log}\left(\theta\left(a\varepsilon^\alpha\right)\right) + \mathrm{Log}(1/\gamma)\right) \implies U(m, \gamma) \leq \varepsilon, \quad (3.1)$$

and also

$$m \geq c' \left(\frac{\nu+\varepsilon}{\varepsilon^2}\right)(d\mathrm{Log}\,(\theta(\nu+\varepsilon)) + \mathrm{Log}(1/\gamma)) \implies U(m,\gamma) \leq \varepsilon. \tag{3.2}$$

3.2 The Realizable Case

Perhaps the most basic and well-studied passive learning algorithm is the method of *empirical risk minimization*. Specifically, for $m \in \mathbb{N}$ and $\mathcal{L} \in (\mathcal{X}\times\mathcal{Y})^m$, define $\mathrm{ERM}(\mathbb{C},\mathcal{L}) = \operatorname*{argmin}_{h\in\mathbb{C}} \mathrm{er}_{\mathcal{L}}(h)$. That is, $\mathrm{ERM}(\mathbb{C},\mathcal{L})$ returns a classifier in $\mathbb{C}$ making the minimum number of mistakes; technically, this really defines an entire family of algorithms, since there may be many classifiers $h \in \mathbb{C}$ with minimal $\mathrm{er}_{\mathcal{L}}(h)$; the results below will hold for every method of breaking such ties. Lemma 3.1 has clear implications for the label complexity of $\mathrm{ERM}(\mathbb{C},\cdot)$ for $\mathcal{P}_{XY}$ in the realizable case (which has $\nu = 0$, and furthermore satisfies Condition 2.3 with $a = \alpha = 1$), as summarized in the following theorem.

Theorem 3.2. The passive learning algorithm $\mathrm{ERM}(\mathbb{C},\cdot)$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ in the realizable case, $\forall \varepsilon,\delta \in (0,1)$,

$$\Lambda(\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim \left(\frac{1}{\varepsilon}\right)(d\mathrm{Log}\,(\theta(\varepsilon)) + \mathrm{Log}\,(1/\delta)).$$

There are other passive learning methods that are known to be sometimes better than $\mathrm{ERM}(\mathbb{C},\cdot)$ in label complexity; for instance, Haussler, Littlestone, and Warmuth [1994] propose a passive learning algorithm that achieves a label complexity Λ that, for $\mathcal{P}_{XY}$ in the realizable case, satisfies

$$\Lambda(\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim \frac{d}{\varepsilon}\ln\left(\frac{1}{\delta}\right),$$

which is sometimes smaller than the label complexity in Theorem 3.2.

However, in some sense, there is not too much room to improve over these results. One basic sense in which this is true is provided by the following lower bound on the *minimax* label complexity, due to Ehrenfeucht, Haussler, Kearns, and Valiant [1989] and Blumer, Ehrenfeucht, Haussler, and Warmuth [1989], which differs from the upper bound of Theorem 3.2 only by a logarithmic factor.

Theorem 3.3. If $|\mathbb{C}| \geq 3$, then for any label complexity Λ achieved by a passive learning algorithm, for any $\varepsilon \in (0, 1/8)$ and $\delta \in (0, 1/100)$, there exists a distribution $\mathcal{P}_{XY}$ in the realizable case for which

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \geq \max\left\{\frac{d-1}{32\varepsilon}, \frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right)\right\}.$$

There is also a somewhat stronger type of lower bound, studied by Antos and Lugosi [1998], who show that for many commonly-used hypothesis classes $\mathbb{C}$, for any label complexity Λ achieved by a passive learning algorithm, there exists a distribution $\mathcal{P}_{XY}$ in the realizable case such that, for every $\delta \in (0, 1)$,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \neq o(1/\varepsilon). \tag{3.3}$$

3.3 The Noisy Case

Lemma 3.1 also has clear implications for the label complexity of $\mathrm{ERM}(\mathbb{C}, \cdot)$ for general $\mathcal{P}_{XY}$. In particular, it implies the following general result.

Theorem 3.4. The passive learning algorithm $\mathrm{ERM}(\mathbb{C}, \cdot)$ achieves a label complexity Λ such that, for any distribution $\mathcal{P}_{XY}$, $\forall \varepsilon, \delta \in (0, 1)$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \left(\frac{\nu + \varepsilon}{\varepsilon^2}\right)\left(d\mathrm{Log}\left(\theta(\nu + \varepsilon)\right) + \mathrm{Log}(1/\delta)\right),$$

and if $\mathcal{P}_{XY}$ satisfies Condition 2.3 with values a and α, then

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim a\left(\frac{1}{\varepsilon}\right)^{2-\alpha}\left(d\mathrm{Log}\left(\theta\left(a\varepsilon^{\alpha}\right)\right) + \mathrm{Log}(1/\delta)\right).$$

Aside from the logarithmic factors, the above result is also known to be nearly minimax-optimal, as reflected by the following lower bound (see e.g., Anthony and Bartlett, 1999, Massart and Nédélec, 2006, Hanneke, 2011, Castro and Nowak, 2008, and Chapter 4 for constructions leading to this result).

Theorem 3.5. There is a universal constant $q \in (0,1)$ such that, if $|\mathbb{C}| \geq 3$, for any label complexity Λ achieved by a passive learning algorithm, for any $\nu \in (0,1/2)$ and sufficiently small $\varepsilon, \delta > 0$, there exists a distribution $\mathcal{P}_{XY}$ for which $\mathrm{er}(f^{\star}) = \nu$ and

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \geq q \left(\frac{\nu + \varepsilon}{\varepsilon^2} \right) (d + \mathrm{Log}(1/\delta)) .$$

Furthermore, for $a \in [2, \infty)$, $\alpha \in (0,1]$, and sufficiently small $\varepsilon, \delta > 0$, there exists a distribution $\mathcal{P}_{XY}$ satisfying Condition 2.3 (in fact, satisfying (2.1) or (2.2), for $\alpha < 1$ or $\alpha = 1$, respectively) with these values of a and α, such that

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \geq qa \left(\frac{1}{\varepsilon} \right)^{2-\alpha} (d + \mathrm{Log}(1/\delta)) .$$

These results will serve as a baseline for comparison when discussing the label complexities achievable by active learning methods below.

4

Lower Bounds on the Label Complexity

Before getting into the development and analysis of active learning algorithms, it will be helpful to set a target for what types of improvements in label complexity we are aiming for. Toward this end, it is good to have general lower bounds, to which we can compare the label complexity upper bounds derived in later sections. In this brief section, we survey some of the known lower bounds on the minimax label complexity of active learning, which hold universally, in the sense that they do not require additional active-learning-specific complexity measures. In Chapter 8, we discuss several other lower bounds that are sometimes tighter, but require the introduction of additional parameters to describe the complexity of the active learning problem. For brevity, we will only provide high-level descriptions of the proofs of some of these results; the reader is referred to the cited original sources for detailed proofs.

4.1 A Lower Bound for the Realizable Case

The most basic lower bound is a classic information-theoretic result of Kulkarni, Mitter, and Tsitsiklis [1993], which in fact holds for any algorithm based on queries that have only two possible answers; since label requests can be answered only as -1 or $+1$, the active learning framework studied here meets this criterion. To state the lower bound, we need to introduce the notion of covering numbers. Specifically, for a given distribution $\mathcal{P}$, the ε-covering number of $\mathbb{C}$ is the smallest integer N such that there exists a set of classifiers $\mathcal{H}$ with $|\mathcal{H}| = N$ for which $\bigcup_{h \in \mathcal{H}} \mathrm{B}(h, \varepsilon) = \mathbb{C}$. We denote the ε-covering number of $\mathbb{C}$ as $\mathcal{N}(\varepsilon, \mathcal{P})$. The lower bound of Kulkarni, Mitter, and Tsitsiklis [1993] is then stated as follows.

Theorem 4.1. For any distribution $\mathcal{P}$, and any label complexity Λ achieved by an active learning algorithm, for any $\varepsilon > 0$, there exists a distribution $\mathcal{P}_{XY}$ in the realizable case with marginal $\mathcal{P}$ over $\mathcal{X}$ such that

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \geq \lceil \log_2 \left((1 - \delta) \mathcal{N}(2\varepsilon, \mathcal{P}) \right) \rceil .$$

The key idea of the proof is to use the fact that there is a set of classifiers $H \subseteq \mathbb{C}$ with $|H| \geq \mathcal{N}(2\varepsilon, \mathcal{P})$ such that any distinct $h, g \in H$ have $\mathcal{P}(x : h(x) \geq g(x)) > 2\varepsilon$ (called a 2ε-packing of $\mathbb{C}$). Then, supposing the target function is chosen at random among H, any learning algorithm that succeeds in producing h of $\mathrm{er}(h) \leq \varepsilon$ effectively *identifies* which $g \in H$ is the target (i.e., $g = \mathrm{argmin}_{f \in H} \mathcal{P}(x : h(x) \neq f(x))$). The average number of bits needed to describe which $g \in H$ is the target (i.e., the entropy) is at least $\log_2(\mathcal{N}(2\varepsilon, \mathcal{P}))$, and since the answers to the queries are binary, this is essentially the source of the lower bound; the factor $(1 - \delta)$ in the logarithm above is included to account for the fact that we allow the algorithm to fail with δ probability.

This lower bound has implications for what the best results we could possibly hope for would look like. In particular, Kulkarni [1989] and Kulkarni, Mitter, and Tsitsiklis [1993] show that if $\mathbb{C}$ has infinite cardinality, then for any sufficiently small $\varepsilon > 0$, there exists $\mathcal{P}$ for which $\log_2(\mathcal{N}(\varepsilon, \mathcal{P})) \gtrsim \max\{d, \log_2(1/\varepsilon)\}$. Furthermore, for many natural classes $\mathbb{C}$, including linear separators [Long, 1995], there are distributions $\mathcal{P}$ for which $\log_2(\mathcal{N}(\varepsilon, \mathcal{P})) \gtrsim d \log(1/\varepsilon)$. This means that, if we choose to express our results in terms of the VC dimension d, we should typically expect label complexity bounds that are at least as large as d and $\log_2(1/\varepsilon)$, and often as large as $d \log(1/\varepsilon)$. Indeed, our upper bounds below will typically contain an explicit factor of the type $d \log(1/\varepsilon)$, in addition to other factors.

4.2 Lower Bounds for the Noisy Cases

The above bound for the realizable case holds for all distributions $\mathcal{P}$, and is expressed in terms of the distribution-dependent quantity $\mathcal{N}(2\varepsilon, \mathcal{P})$. There are also known lower bounds for the noisy case that are often larger than the above, but the distributions they hold for have stronger requirements (needed for the construction of the hard scenarios leading to the lower bounds). In the noisy case, we are interested in showing that, for any label complexity Λ achieved by an active learning algorithm, there exists a distribution $\mathcal{P}_{XY}$ satisfying some specified noise conditions, for which $\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY})$ is at least a certain size.

The lower bounds for active learning with noise can largely be traced to the work of Kääriäinen [2006], who proved a lower bound of order ν^2/ε^2 holding for essentially any nontrivial hypothesis class, for a certain type of distribution $\mathcal{P}$ that he constructed. Beygelzimer, Dasgupta, and Langford [2009] later strengthened this to $d\nu^2/\varepsilon^2$, essentially by constructing $d-1$ independent problems of the same type constructed by Kääriäinen [2006]. By slightly modifying the construction of Kääriäinen [2006], Hanneke [2011] showed that (for nontrivial $\mathbb{C}$), there exist distributions $\mathcal{P}_{XY}$ satisfying Condition 2.3 for which $\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \gtrsim \varepsilon^{2\alpha-2}$; this generalized an earlier result of Castro and Nowak [2008] showing a similar result for the specific class of threshold classifiers (Example 1).

Each of these lower bounds has been largely based on a basic information-theoretic lower bound on the number of times one must flip a given biased coin in order to confidently decide whether the coin is biased toward heads or tails. Results of this type originate in the work of Wald [1945]. The following variant is taken from Anthony and Bartlett [1999].

Lemma 4.2. Fix any $\gamma \in (0,1)$, $\delta \in (0,1/4)$, and $n \in \mathbb{N}$, and let $p_0 = 1/2 - \gamma/2$ and $p_1 = 1/2 + \gamma/2$. Fix any function $\hat{t} : \{0,1\}^n \to \{0,1\}$

(possibly randomized). If

$$n < 2 \left\lfloor \frac{1-\gamma^2}{2\gamma^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right) \right\rfloor,$$

then for $t \sim \text{Bernoulli}(1/2)$, and $B_1, \ldots, B_n$ conditionally independent (given t) Bernoulli(p_t) random variables (with t and $B_1, \ldots, B_n$ all independent of $\hat{t}$), with probability greater than δ, $\hat{t}(B_1, \ldots, B_n) \neq t$.

The following theorem combines the techniques of Beygelzimer, Dasgupta, and Langford [2009] and Hanneke [2011]; in particular, this result is slightly stronger than those appearing in the published literature (for Condition 2.3).

Theorem 4.3. There exists a universal constant $q \in (0, \infty)$ such that, if $|\mathbb{C}| \geq 3$, then for any label complexity Λ achieved by an active learning algorithm, for any $\nu \in (0, 1/2)$ and sufficiently small $\varepsilon, \delta > 0$, there exists a distribution $\mathcal{P}_{XY}$ with $\text{er}(f^\star) = \nu$ such that

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \geq q \left(\frac{\nu^2}{\varepsilon^2}\right)(d + \text{Log}(1/\delta)). \tag{4.1}$$

Furthermore, for any $a \in [4, \infty)$, $\alpha \in (0, 1]$, and sufficiently small $\varepsilon, \delta > 0$, there exists a distribution $\mathcal{P}_{XY}$ satisfying Condition 2.3 (in fact, satisfying (2.1) or (2.2), for $\alpha < 1$ or $\alpha = 1$, respectively) with these values of a and α, such that

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \geq q a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha}(d + \text{Log}(1/\delta)). \tag{4.2}$$

Proof. The proof is in two parts, corresponding to the d term and the $\text{Log}(1/\delta)$ term, respectively, from each of these lower bounds.

If $d = 1$, the d term is redundant, so we may skip this step. Otherwise, suppose $d \geq 2$. Let $\{x_0, \ldots, x_{d-1}\}$ denote a set of d points shattered by $\mathbb{C}$. Define the distribution $\mathcal{P}$ as follows. For each $i \in \{1, \ldots, d-1\}$, let $\mathcal{P}(\{x_i\}) = \beta/(d-1)$, for a value $\beta \in [24\varepsilon, 1)$ to be determined below. Then let $\mathcal{P}(\{x_0\}) = 1 - \beta$. Also fix a value $\gamma = 24\varepsilon/\beta$, and let $p_0 = 1/2 - \gamma/2$ and $p_1 = 1/2 + \gamma/2$. Fix the learning algorithm achieving label complexity Λ, and for any $n \in \mathbb{N}$ and

$t = \{t_i\}_{i=1}^{d-1} \in \{0,1\}^{d-1}$, let $\hat{h}_{nt}$ be the classifier the active learning algorithm would produce when given budget n and run under the distribution $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(t)}$ that has marginal $\mathcal{P}$ on $\mathcal{X}$ and has $\eta(x_0) = 1$ and $\eta(x_i) = p_{t_i}$ for each $i \in \{1, \ldots, d-1\}$. Furthermore, let m_{tj} denote the index of the j^{th} label (namely, $Y_{m_{tj}}$) requested by the active learning algorithm, when run under distribution $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(t)}$. Note that, since the Y_j values are conditionally independent given the X_j values, we may assume (without loss of generality) that for each j, the index m_{tj} is minimal such that every $m' < m_{tj}$ with $X_{m'} = X_{m_{tj}}$ already has $m' \in \{m_{tj'}\}_{j'=1}^{j-1}$ (i.e., $X_{m_{tj}}$ is the earliest instance in the sequence at that particular location for which the label has not yet been requested). More precisely, for any algorithm for which this is not the case, there is another for which it is the case with a distributionally equivalent output $\hat{h}_{nt}$. Also, for each $i \in \{0, \ldots, d-1\}$ and $j \in \mathbb{N}$, let $Y_{ij} = Y_k$ for $k \in \mathbb{N}$ such that $X_k = x_i$ and $|\{k' < k : X_{k'} = x_i\}| = j - 1$ (assuming such a k exists): that is, Y_{ij} is the j^{th} label in the sequence for which the corresponding $X_m = x_i$.

Now let $m = 2\left\lfloor \frac{1-\gamma^2}{2\gamma^2} \ln\left(\frac{9}{8}\right)\right\rfloor$. For each $i \in \{1, \ldots, d-1\}$, let $\hat{t}_{nti} = (\hat{h}_{nt}(x_i) + 1)/2 \in \{0,1\}$ if $|\{m_{tj} : j \le n, X_{m_{tj}} = x_i\}| < m$, and $\hat{t}_{nti} = 0$ otherwise. Now consider taking $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(\mathbf{t})}$ where $\mathbf{t} \sim \text{Uniform}(\{0,1\}^{d-1})$; that is, $\mathbf{t} = \{\mathbf{t}_i\}_{i=1}^{d-1}$ are i.i.d. Bernoulli(1/2), and the data $\mathcal{Z}$ are conditionally (given $\mathbf{t}$) i.i.d. $\mathcal{P}_{XY}^{(\mathbf{t})}$. For each $i \in \{1, \ldots, d-1\}$, $\hat{t}_{n\mathbf{t}i}$ is a function of $\{(Y_{ij} + 1)/2 : j < m\}$ (along with other independent random variables: namely, $\mathcal{Z}_{\mathbf{X}}$, $\{Y_{i'j} : i' \neq i, j \in \mathbb{N}\}$, and any independent randomness internal to the algorithm), and these $(Y_{ij} + 1)/2$ values are (conditionally, given $\mathbf{t}_i$) i.i.d. Bernoulli($p_{\mathbf{t}_i}$) random variables. Thus, by Lemma 4.2, with probability greater than 1/3, $\hat{t}_{n\mathbf{t}i} \neq \mathbf{t}_i$. In particular, if we suppose $n < m(d-1)/12$, there must exist at least $(d-1)/2$ of the points $x_i \in \{x_1, \ldots, x_{d-1}\}$ for which, with probability at least 5/6, $|\{m_{\mathbf{t}j} : j \le n, X_{m_{\mathbf{t}j}} = x_i\}| < m$: that is, the active learning algorithm requests fewer than m labels Y_j with $X_j = x_i$ (otherwise, the expected number of requested labels would exceed n). Combined with the above guarantee on $\hat{t}_{n\mathbf{t}i}$ supplied by Lemma 4.2, this means at least $(d-1)/2$ of these x_i points have, with probability greater than $1/3 - 1/6 = 1/6$, $(\hat{h}_{n\mathbf{t}}(x_i) + 1)/2 = \hat{t}_{n\mathbf{t}i} \neq \mathbf{t}_i = (f^{\star}(x_i) + 1)/2$, so that $\hat{h}_{n\mathbf{t}}(x_i) \neq f^\star(x_i)$. Thus, by linearity of the expectation, the expected number of values $i \in \{1,\dots,d-1\}$ for which $\hat{h}_{n\mathbf{t}}(x_i) \neq f^\star(x_i)$ is greater than $(d-1)/12$. Since the number of such i can never be more than $d-1$, this implies that, with probability greater than $1/24$, the number of $i \in \{1,\dots,d-1\}$ with $\hat{h}_{n\mathbf{t}}(x_i) \neq f^\star(x_i)$ is greater than $(d-1)/24$. Thus, with probability greater than $1/24$,

$$\mathrm{er}(\hat{h}_{n\mathbf{t}}) - \mathrm{er}(f^\star) \geq \sum_{i=1}^{d-1} \gamma(\beta/(d-1))\mathbb{1}[\hat{h}_{n\mathbf{t}}(x_i) \neq f^\star(x_i)] > \gamma\beta/24 = \varepsilon. \tag{4.3}$$

By the law of total probability, and the fact that $\max(\cdot)$ upper bounds average$(\cdot)$, there exists a $t \in \{0,1\}^{d-1}$ such that, given that $\mathbf{t} = t$, the conditional probability that (4.3) holds is greater than $1/24$. In other words, for this choice of t, taking $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(t)}$ guarantees that, with probability greater than $1/24$, $\mathrm{er}(\hat{h}_{nt}) - \mathrm{er}(f^\star) > \varepsilon$. Thus, for $\delta \leq 1/24$, we must have $\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) > n$.

We can prove the $\mathrm{Log}(1/\delta)$ term very similarly. Since $|\mathbb{C}| \geq 3$, there must exist $\tilde{x}_0, \tilde{x}_1 \in \mathcal{X}$ and $h_0, h_1 \in \mathbb{C}$ such that $h_0(\tilde{x}_0) = h_1(\tilde{x}_1)$ while $h_0(\tilde{x}_1) \neq h_1(\tilde{x}_1)$. Now specify the distribution $\mathcal{P}$ by letting $\mathcal{P}(\{\tilde{x}_1\}) = \beta$ and $\mathcal{P}(\{\tilde{x}_0\}) = 1-\beta$, for a value $\beta \in [24\varepsilon, 1)$ to be specified below. Let γ, p_0, and p_1 be as above. However, this time, for $s \in \{0,1\}$, we say $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(s)}$ when the marginal distribution over $\mathcal{X}$ is $\mathcal{P}$ and when $\eta(\tilde{x}_0) = (h_0(\tilde{x}_0)+1)/2$ and $\eta(\tilde{x}_1) = p_s$, and we let $\hat{h}_{ns}$ denote the classifier returned by the algorithm under these conditions. Again, without loss of generality, we can suppose that every time the algorithm requests the label Y_m for some $X_m = \tilde{x}_1$, the value m is the smallest for which $X_m = \tilde{x}_1$ and the label Y_m has not yet been requested. Furthermore, since the labels Y_m with $X_m = \tilde{x}_0$ have $Y_m = h_0(\tilde{x}_0)$ (with probability 1), we may regard $\hat{h}_{ns}$ as a function of $\{(\tilde{Y}_{1m}+1)/2\}_{m=1}^n$ (along with other independent random variables namely, $\mathcal{Z}_{\mathbf{X}}$), where $\tilde{Y}_{1m} = Y_k$ for the value $k \in \mathbb{N}$ such that $X_k = \tilde{x}_1$ and $|\{k' < k : X_{k'} = \tilde{x}_1\}| = m-1$ (assuming such a k exists, which in our case happens with probability one). Since $\{(\tilde{Y}_{1m}+1)/2\}_{m=1}^n$ is a sequence of n independent Bernoulli(p_s) random variables, Lemma 4.2 (combined with the law of total probability and the fact that $\max(\cdot)$ upper bounds average$(\cdot)$) implies that, if $n < 2\left\lfloor \frac{1-\gamma^2}{2\gamma^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor$, there is a choice of $s \in \{0,1\}$ for which, with probability greater than δ, the value $\hat{s}_{ns} = (\hat{h}_{ns}(\tilde{x}_1)+1)/2$ has $\hat{s}_{ns} \neq s = (f^\star(\tilde{x}_1)+1)/2$, so that $\hat{h}_{ns}(\tilde{x}_1) \neq f^\star(\tilde{x}_1)$. On this event we have $\mathrm{er}(\hat{h}_{ns}) - \mathrm{er}(f^\star) \geq \gamma\beta = 24\varepsilon$. In particular, this implies $\mathcal{P}_{XY} = \mathcal{P}_{XY}^{(s)}$ has $\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) > n$.

The above two analyses hold for an arbitrary choice of $\beta \in [24\varepsilon, 1)$, and result in a combined lower bound $\gtrsim (\beta^2/\varepsilon^2)(d \vee \mathrm{Log}(1/\delta)) \gtrsim (\beta^2/\varepsilon^2)(d + \mathrm{Log}(1/\delta))$ when $\delta \in (0, 1/24]$ and $\varepsilon \in (0,1)$ is sufficiently small relative to β. To obtain the two lower bounds in the theorem statement, we need to set β so that the distribution $\mathcal{P}_{XY}$ in each of the above constructions satisfies the respective conditions of these two results. Note that the distributions $\mathcal{P}_{XY}$ constructed above satisfy $\mathrm{er}(f^\star) = \beta(1/2 - \gamma/2) = \beta/2 - 12\varepsilon$. Thus, to obtain (4.1) for a given value of ν, we can take $\beta = 2(\nu + 12\varepsilon)$, which guarantees $\mathrm{er}(f^\star) = \nu$ (as required). For (4.2), the value ν is considered a free variable, and we need only set β so that the above $\mathcal{P}_{XY}$ constructions satisfy Condition 2.3 for the given a and α values. In the case of $\alpha = 1$, we can take $\beta = 24a\varepsilon$, in which case $\gamma = 1/a$, and therefore $\mathcal{P}(x : |\eta(x) - 1/2| < 1/(2a)) = \mathcal{P}(x : |\eta(x) - 1/2| < \gamma/2) = 0$, so that the above distributions $\mathcal{P}_{XY}$ satisfy (2.2), and hence Condition 2.3. For the remaining cases of $\alpha \in (0,1)$, note that the set of $x \in \mathcal{X}$ with $|\eta(x) - 1/2| \leq \gamma/2$ has probability β. So setting $\beta = (1-\alpha)^{1-\alpha}(2\alpha)^\alpha a 12^\alpha \varepsilon^\alpha$ satisfies $\beta = a'(12\varepsilon/\beta)^{\alpha/(1-\alpha)} = a'(\gamma/2)^{\alpha/(1-\alpha)}$, where $a' = (1-\alpha)(2\alpha)^{\alpha/(1-\alpha)}a^{1/(1-\alpha)}$. Furthermore, for any $t \in (\gamma/2, 1/2)$, $\mathcal{P}(x : |\eta(x) - 1/2| \leq t) = \mathcal{P}(x : |\eta(x) - 1/2| \leq \gamma/2) \leq a'(\gamma/2)^{\alpha/(1-\alpha)} \leq a't^{\alpha/(1-\alpha)}$. Also, note that any $t \geq 1/2$ has $\mathcal{P}(x : |\eta(x) - 1/2| \leq t) = 1 \leq (1-\alpha)\alpha^{\alpha/(1-\alpha)}2^{1/(1-\alpha)} \leq a'(1/2)^{\alpha/(1-\alpha)} \leq a't^{\alpha/(1-\alpha)}$. On the other hand, any $t < \gamma/2$ has $\mathcal{P}(x : |\eta(x) - 1/2| \leq t) = 0 \leq a't^{\alpha/(1-\alpha)}$. Therefore, $\mathcal{P}_{XY}$ satisfies (2.1) for the values a' and α, and hence Condition 2.3 as well. Thus, these choices of β suffice for the two respective results, as long as ε is small enough in each case to have, for instance, $\beta \in [48\varepsilon, 1)$. □

The general implication of Theorem 4.3 is that we should expect our label complexity results below to be at least as large as $(\nu^2/\varepsilon^2)(d + \mathrm{Log}(1/\delta))$ when expressed in terms of ν, and at least as large as $a^2(1/\varepsilon)^{2-2\alpha}(d + \mathrm{Log}(1/\delta))$ when expressed in terms of a and

α. In Chapter 5, we will find that a very simple technique is able to achieve these lower bounds (up to logarithmic factors) under certain well-understood conditions: namely, when the disagreement coefficient is bounded by a constant. Later, Chapter 8 discusses more-involved techniques that nearly achieve the lower bound under more general conditions.

5

Disagreement-Based Active Learning

This section discusses a technique for the design of active learning algorithms, and the analysis thereof, based on a very simple principle: never request the label Y_i of a point X_i if we can derive the value of $f^\star(X_i)$ from information already available. In some sense, this represents the *least* we should expect from a reasonable active learning algorithm. More explicitly, algorithms based on this principle maintain a set of data-dependent constraints satisfied by $f^\star$ with high probability (involving empirical error rates on the previously-queried data points), and they request the label Y_i of the next data point X_i in the sequence if and only if there are at least two classifiers in $\mathbb{C}$ satisfying these constraints but disagreeing on the label of X_i. Variants of this general idea were first discussed in the pioneering works of Cohn, Atlas, and Ladner [1994] and Balcan, Beygelzimer, and Langford [2006], and have since amassed a substantial literature. As we will see in later chapters, this technique does not always yield optimal label complexities. However, due to its elegance, and other favorable properties, including robustness to noise, this has become one of the most commonly-studied techniques in the theory of active learning. It is therefore worth forming a thorough understanding of this technique and the label complexities it can achieve.

5.1 The Realizable Case: CAL

Perhaps one of the earliest and most elegant general-purpose active learning algorithms designed for the realizable case was proposed by Cohn, Atlas, and Ladner [1994], and is now typically referred to as *CAL* after these authors. The algorithm is specified as follows.

Algorithm: $\mathbf{CAL}(n)$
0. $m \leftarrow 0$, $Q \leftarrow \{\}$
1. While $\lvert Q\rvert < n$ and $m < 2^n$
2. $\quad m \leftarrow m + 1$
3. $\quad$ If $\forall y \in \mathcal{Y}, \exists h \in \mathbb{C}$ s.t. $\mathrm{er}_{Q\cup\{(X_m,y)\}}(h) = 0$
4. $\qquad$ Request label Y_m; let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$
5. Return any $\hat{h} \in \mathbb{C}$ with $\mathrm{er}_Q(\hat{h}) = 0$

Executing this algorithm requires us to solve a sequence of constraint-satisfaction problems (Step 3), and to find an explicit solution to such a constraint-satisfaction problem in Step 5. The condition in Step 3 checks for whether there exist two classifiers $h, g \in \mathbb{C}$ that are correct on all of the observed labels so far ($\mathrm{er}_Q(h) = \mathrm{er}_Q(g) = 0$), and yet disagree on the label of the new point ($h(X_m) \neq g(X_m)$). The fact that we suppose $|\mathcal{Y}| = 2$ allows us to express this condition in the somewhat more concise form stated as Step 3. If the condition is satisfied, the algorithm requests the label in Step 4.

Although the above pseudo-code represents a fairly good description of the types of computations required to execute each step of CAL (namely, solving various constraint-satisfaction problems), for the purpose of simplifying the analysis, this algorithm is often expressed in a form that makes these steps more *implicit*, so as to give an explicit name to the set of classifiers $h \in \mathbb{C}$ satisfying the constraint $\mathrm{er}_Q(h) = 0$. This set is typically referred to as the *version space*. Specifically, the following is an equivalent form of CAL.

Algorithm: $\mathbf{CAL}(n)$
0. $m \leftarrow 0$, $t \leftarrow 0$, $V \leftarrow \mathbb{C}$
1. While $t < n$ and $m < 2^n$
2. $\quad m \leftarrow m + 1$
3. $\quad$ If $X_m \in \mathrm{DIS}(V)$
4. $\quad\quad$ Request label Y_m; let $V \leftarrow \{h \in V : h(X_m) = Y_m\}$, $t \leftarrow t + 1$
5. Return any $\hat{h} \in V$

Expressed in this form, the analysis of the label complexity achieved by CAL then focuses on bounding the number of label requests sufficient to guarantee every h in the version space V has $\mathrm{er}(h) \leq \varepsilon$ with probability at least $1 - \delta$. In particular, note that in the realizable case, every $Y_m = f^\star(X_m)$, so that the update in Step 4 guarantees $f^\star \in V$ is maintained as an invariant, and furthermore that after the update in Step 4, every $h \in V$ must have $h(X_m) = Y_m = f^\star(X_m)$. The algorithm only refrains from requesting a label Y_m if every $h \in V$ classifies X_m the same: namely, as $f^\star(X_m)$, since $f^\star$ is among them. Hence, by induction, after processing unlabeled data points $X_1, \ldots, X_m$, the version space V is the set of classifiers $h \in \mathbb{C}$ that agree with $f^\star$ on all of $X_1, \ldots, X_m$: said another way, since $f^\star(X_i) = Y_i$ for each $i \leq m$, at the end of round m we can express $V = \{h \in \mathbb{C} : \mathrm{er}_m(h) = 0\} = V_m^\star$. Thus, the classifier $\hat{h}$ returned by $\mathrm{CAL}(n)$ is equivalent to that returned by $\mathrm{ERM}(\mathbb{C}, \mathcal{Z}_m)$, for the largest value of m obtained in the algorithm. Therefore, Theorem 3.2 already describes the size of m sufficient to guarantee $\mathrm{er}(\hat{h}) \leq \varepsilon$ with high probability, and the analysis of the label complexity of CAL then reduces to bounding the number of label requests the algorithm would make among that many unlabeled data points, which is characterized by the rate of collapse of the value $\mathcal{P}(\mathrm{DIS}(V))$ as the algorithm proceeds. This in turn can be bounded in terms of the disagreement coefficient, in combination with Lemma 3.1. Formally, we have the following theorem, originally due to Hanneke [2011] (though the proof below is somewhat different from the original).

Theorem 5.1. CAL achieves a label complexity Λ such that, for $\mathcal{P}_{XY}$

in the realizable case, $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \theta(\varepsilon)\left(d\mathrm{Log}(\theta(\varepsilon)) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right)\mathrm{Log}(1/\varepsilon). \quad (5.1)$$

Proof. Fix any $\varepsilon, \delta \in (0,1)$, and consider running CAL with budget argument $n \in \mathbb{N}$ satisfying

$$n \geq \log_2(2/\delta) + 8ec'\theta(\varepsilon)\left(d\mathrm{Log}(\theta(\varepsilon)) + 2\mathrm{Log}(2\log_2(4/\varepsilon)/\delta)\right)\log_2(2/\varepsilon).$$

Let $M \subseteq \{0, \ldots, 2^n\}$ denote the set of values of m obtained during the execution. For each $m \in M$, let V_m denote the value of V upon reaching Step 1 with that value of m. As discussed above, on an event E of probability 1, every $m \in \mathbb{N}$ has $Y_m = f^\star(X_m)$, which implies $\forall m \in M, f^\star \in V_m = V_m^\star = \{h \in \mathbb{C} : \mathrm{er}_m(h) = 0\}$ by induction.

Now let $i_\varepsilon = \lceil \log_2(1/\varepsilon) \rceil$ and define $I = \{0, \ldots, i_\varepsilon\}$. For $i \in I$, let $\varepsilon_i = 2^{-i}$; furthermore, let $m_0 = 0$, and for c' as in (3.2), for each $i \in I \setminus \{0\}$, define

$$m_i = \left\lceil c'\left(\frac{1}{\varepsilon_i}\right)\left(d\mathrm{Log}(\theta(\varepsilon_i)) + \mathrm{Log}\left(\frac{2(2+i_\varepsilon-i)^2}{\delta}\right)\right)\right\rceil.$$

Lemma 3.1, (3.2), and a union bound imply that, on an event E_δ of probability at least $1 - \sum_{i=1}^{i_\varepsilon} \frac{\delta}{2(2+i_\varepsilon-i)^2} > 1 - \delta/2$, every $i \in I$ has

$$\sup_{h \in V_{m_i}^\star} \mathrm{er}(h) \leq \varepsilon_i. \quad (5.2)$$

Now note that, on event E, the total number of label requests made by CAL while $m \leq m_{i_\varepsilon}$ is exactly

$$\sum_{m=1}^{\min\{m_{i_\varepsilon}, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}(X_m) = \sum_{m=1}^{\min\{m_{i_\varepsilon}, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{m-1}^\star)}(X_m),$$

which is at most

$$\sum_{m=1}^{m_{i_\varepsilon}} \mathbb{1}_{\mathrm{DIS}(V_{m-1}^\star)}(X_m) = \sum_{i \in I \setminus \{0\}} \sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}(V_{m-1}^\star)}(X_m). \quad (5.3)$$

By monotonicity of $V_m^\star$ in m, any $i \in I \setminus \{0\}$ and $m \in \{m_{i-1}+1, \ldots, m_i\}$ have $\mathrm{DIS}(V_{m-1}^\star) \subseteq \mathrm{DIS}(V_{m_{i-1}}^\star)$, and (5.2) implies that on event E_δ,

$V^{\star}_{m_{i-1}} \subseteq \mathrm{B}(f^{\star}, \varepsilon_{i-1})$, so that $\mathrm{DIS}(V^{\star}_{m_{i-1}}) \subseteq \mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1}))$. Thus, (5.3) is at most

$$\sum_{i \in I \setminus \{0\}} \sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1}))}(X_m). \tag{5.4}$$

This is a sum of m_{i_ε} independent Bernoulli random variables, with expected value

$$\sum_{i \in I \setminus \{0\}} (m_i - m_{i-1}) \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1}))).$$

Thus, a Chernoff bound implies that, on an event E'_δ of probability at least $1 - \delta/2$, (5.4) is at most

$$\log_2(2/\delta) + 2e \sum_{i \in I \setminus \{0\}} (m_i - m_{i-1}) \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1}))), \tag{5.5}$$

By definition of the disagreement coefficient, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1}))) \leq \theta(\varepsilon_{i-1})\varepsilon_{i-1}$, and combining this with the definition of m_i, we have that for $i \in I \setminus \{0\}$, $m_i \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, \varepsilon_{i-1})))$ is at most

$$4c'\theta(\varepsilon_{i-1}) \left(d\mathrm{Log}(\theta(\varepsilon_i)) + \mathrm{Log}\left(\frac{2(2 + i_\varepsilon - i)^2}{\delta} \right) \right). \tag{5.6}$$

Thus, since $\theta(\varepsilon_{i-1}) \leq \theta(\varepsilon_i) \leq \theta(\varepsilon)$ for $i \in I \setminus \{0\}$, (5.5) is less than

$$\log_2(2/\delta) + 8ec'\theta(\varepsilon) \left(d\mathrm{Log}(\theta(\varepsilon)) + 2\mathrm{Log}(2\log_2(4/\varepsilon)/\delta) \right) \log_2(2/\varepsilon) \leq n. \tag{5.7}$$

In particular, we have proven that on event $E \cap E_\delta \cap E'_\delta$, $\sup_{h \in V^{\star}_{m_{i_\varepsilon}}} \mathrm{er}(h) \leq \varepsilon_{i_\varepsilon} \leq \varepsilon$, and the number of label requests made by CAL while $m \leq m_{i_\varepsilon}$ is less than n; since $2^n > m_{i_\varepsilon}$, this means we must have $\max M \geq m_{i_\varepsilon}$, so that $\hat{h} \in V^{\star}_{m_{i_\varepsilon}}$, and thus $\mathrm{er}(\hat{h}) \leq \varepsilon$.

Noting that $E \cap E_\delta \cap E'_\delta$ has probability at least $1 - \delta$ (by a union bound), and that

$$\begin{aligned} &\log_2(2/\delta) + 8ec'\theta(\varepsilon) \left(d\mathrm{Log}(\theta(\varepsilon)) + 2\mathrm{Log}(2\log_2(4/\varepsilon)/\delta) \right) \log_2(2/\varepsilon) \\ &\qquad \lesssim \theta(\varepsilon) \left(d\mathrm{Log}(\theta(\varepsilon)) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)/\delta) \right) \mathrm{Log}(1/\varepsilon), \end{aligned}$$

completes the proof. $\square$

In general, the asymptotic dependence on ε in the bound of Theorem 5.1 is

$$O\left(\theta(\varepsilon)\mathrm{Log}(1/\varepsilon)\mathrm{Log}\Big(\theta(\varepsilon)\mathrm{Log}(1/\varepsilon)\Big)\right).$$

This is particularly interesting when $\theta(\varepsilon) = O(1)$ (equivalently, $\theta(0) < \infty$), especially in comparison to the passive learning label complexity, which (as discussed in Chapter 3) is typically $\Omega(1/\varepsilon)$; see Chapter 7 for several interesting examples of $\mathbb{C}$ and $\mathcal{P}$ for which $\theta(\varepsilon) = O(1)$, as well as general sufficient conditions for this that apply to a broad family of learning problems. In the case of $\theta(\varepsilon) = \Theta(\varepsilon^{-\kappa})$ for some $\kappa \in (0, 1]$, one can improve the bound in Theorem 5.1 by a more careful treatment of the summation that results from plugging (5.6) into (5.5); specifically, this results in a bound $\lesssim \theta(\varepsilon)(d\mathrm{Log}(\theta(\varepsilon)) + \mathrm{Log}(c_\kappa/\delta))c_\kappa$, where c_κ is a κ-dependent constant, so that the asymptotic dependence on ε becomes $O(\theta(\varepsilon)\mathrm{Log}(\theta(\varepsilon)))$.

Aside from the logarithmic factors, one can show the bound of Theorem 5.1 often represents a fairly tight analysis of the label complexity of CAL, especially the asymptotic dependence on ε described by the leading $\theta(\varepsilon)$ factor. The proof of Theorem 5.1 centered on bounding the number of label requests among the first m unlabeled data points, for a choice of $m = \tilde{O}(1/\varepsilon)$ based on the label complexity of $\mathrm{ERM}(\mathbb{C}, \cdot)$. Hanneke [2012] shows that the leading factor of $\theta(\varepsilon)$ also arises in *lower bounds* on the number of labels requested among $1/\varepsilon$ data points. Furthermore, another possible route to bounding the label complexity of CAL (taken by Hanneke, 2011) is to directly bound $\mathcal{P}(\mathrm{DIS}(V))$ as a function of the number of labels requested by the algorithm so far. Hanneke [2012] additionally shows that this factor of $\theta(\varepsilon)$ also arises in a lower bound on the number of label requests the algorithm must make to achieve a certain value of $\mathcal{P}(\mathrm{DIS}(V))$. Formally, for $n, m \in \mathbb{N}$, let $N(m) = \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(V_{i-1}^\star)}(X_i)$, representing the number of labels CAL would request (in the realizable case) among the first m unlabeled data points (assuming it does not halt first), and let $M(n) = \min\{k \in \mathbb{N} : N(k) = n\} \cup \{\infty\}$, representing the number of unlabeled data points CAL would process up to its n^{th} label request (assuming a budget of at least n). We have the following theorems from Hanneke [2012], along with brief sketches to give the highlights of the proofs; the interested reader is referred to the work of Hanneke [2012] for the full proofs.

Theorem 5.2. For any $m \in \mathbb{N} \cup \{0\}$ and $r \in (0,1)$,

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_m^\star\right)\right)\right] \geq (1-r)^m \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r))).$$

Furthermore, this implies that for any $\varepsilon \in (0,1)$,

$$\mathbb{E}\left[N\left(\lceil 1/\varepsilon \rceil\right)\right] \geq \theta(\varepsilon)/2.$$

Proof Sketch. If $x \in \mathrm{DIS}(\mathrm{B}(f^\star, r))$, then there is a classifier $h_x \in \mathbb{C}$ with $\mathcal{P}(x' : h_x(x') \neq f^\star(x')) \leq r$ for which $h_x(x) \neq f^\star(x)$. But then the probability that $h_x(X_i) = f^\star(X_i)$ is at least $(1-r)$, so that the probability $x \in \mathrm{DIS}\left(V_m^\star\right)$ is at least $(1-r)^m$. Since this holds for every $x \in \mathrm{DIS}(\mathrm{B}(f^\star, r))$, we have that for $X \sim \mathcal{P}$ independent of $\mathcal{Z}_{\mathbf{X}}$, the probability $X \in \mathrm{DIS}\left(V_m^\star\right)$ is at least $(1-r)^m \mathbb{P}(X \in \mathrm{DIS}\left(\mathrm{B}(f^\star, r)\right))$. The noted implication follows by summing the resulting geometric series lower bounding $\mathbb{E}[N(\lceil 1/r \rceil)]$, and maximizing over $r > \varepsilon$. □

Theorem 5.3. For any $n \in \mathbb{N}$ and $r \in (0,1)$,

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star\right)\right)\right] \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r))) - nr.$$

Furthermore, this implies that for any $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$,

$$n \leq \theta(\varepsilon)/2 \implies \mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star\right)\right)\right] \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, \varepsilon)))/2.$$

Proof Sketch. Note that for any $i \leq n$, the point $X_{M(i)}$ is (conditionally given $V_{M(i-1)}^\star$) a random sample from the conditional distribution of $X \sim \mathcal{P}$ given $X \in \mathrm{DIS}\left(V_{M(i-1)}^\star\right)$. Thus, for $x \in \mathrm{DIS}\left(V_{M(i-1)}^\star \cap \mathrm{B}(f^\star, r)\right)$ and $h_x \in V_{M(i-1)}^\star \cap \mathrm{B}(f^\star, r)$ with $h_x(x) \neq f^\star(x)$, the conditional probability (given $V_{M(i-1)}^\star$) that $h_x \notin V_{M(i)}^\star$ is at most $r/\mathcal{P}\left(\mathrm{DIS}\left(V_{M(i-1)}^\star\right)\right) \leq r/\mathcal{P}\left(\mathrm{DIS}\left(V_{M(i-1)}^\star \cap \mathrm{B}(f^\star, r)\right)\right)$. This implies that, for $X \sim \mathcal{P}$ independent of $\mathcal{Z}_{\mathbf{X}}$, the probability $X \in \mathrm{DIS}\left(V_{M(i-1)}^\star \cap \mathrm{B}(f^\star, r)\right) \setminus \mathrm{DIS}\left(V_{M(i)}^\star \cap \mathrm{B}(f^\star, r)\right)$ is at most r; since this condition is satisfied for *some* $i \leq n$ anytime $X \in \mathrm{DIS}\left(\mathrm{B}(f^\star, r)\right) \setminus \mathrm{DIS}\left(V_{M(n)}^\star\right)$, the result follows by a union bound. Plugging in any $r \in [\varepsilon, 1)$ and $n \leq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r)))/(2r)$ gives $\mathbb{E}[\mathcal{P}(\mathrm{DIS}(V_{M(n)}^\star))] \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r)))/2 \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, \varepsilon)))/2$, and the noted implication then holds by maximizing over $r \geq \varepsilon$. □

5.1.1 CAL as a Selective Sampling Algorithm

CAL can be expressed as a particular type of active learning algorithm, known as a *selective sampling* algorithm, which visits each unlabeled data point X_m in sequence, and for each m, makes a decision on whether or not to request the label Y_m based only on the previously-observed X_i values ($i \leq m$) and corresponding requested labels, and never changes this decision once made. More formally, a selective sampling algorithm is represented as two sequences of random variables – a sequence $\{q_m\}_{m=1}^{\infty}$ of values in $\{0,1\}$, and a sequence $\{\hat{h}_m\}_{m=0}^{\infty}$ of classifiers – with the characteristic that $\forall m \in \mathbb{N} \cup \{0\}$, $\hat{h}_m$ is conditionally independent of $\mathcal{Z}$ given $\{(X_i, q_i Y_i)\}_{i=1}^{m}$, and $\forall m \in \mathbb{N}$, q_m is conditionally independent of $\mathcal{Z}$ given $\{(X_i, q_i Y_i)\}_{i=1}^{m-1}$ and X_m. These q_m random variables indicate whether or not the algorithm requests the label Y_m.

Clearly any selective sampling algorithm can be made into an active learning algorithm in the sense studied here, simply by returning the classifier $\hat{h}_m$, where m is the smallest index with $\sum_{i=1}^{m} q_i = n$, given a budget n on the number of label requests; this is one way to view the CAL active learning algorithm stated above. It is not hard to see that any selective sampling algorithm that, combined with this stopping criterion, leads to a label complexity with improved asymptotic dependence on ε compared to the passive learning algorithm that returns $\hat{h}_n$ must have $\sum_{t=1}^{m} q_i = o(m)$. Furthermore, the recent work of Hanneke [2012] shows that, in the realizable case, this sublinearity condition is also *sufficient* for CAL to achieve such an improved dependence on ε, compared to the label complexities achievable by the entire family of $\mathrm{ERM}(\mathbb{C}, \cdot)$ passive learning algorithms.

Hanneke [2012] further shows that this sublinearity condition (namely, $\mathbb{E}[N(m)] = o(m)$) holds for CAL if and only if $\theta(\varepsilon) = o(1/\varepsilon)$ (see Lemma 7.12 below for discussion of this latter condition). The "only if" half is supplied by Theorem 5.2 (taking $\varepsilon = 1/m$). The "if" half is given by Lemma 3.1, which implies that (in the realizable case), for $r_i = d\mathrm{Log}(i)/i$, with probability $\geq 1 - r_i$, $\sup_{h \in V_i^\star} \mathrm{er}(h) \lesssim r_i$, so that $\mathbb{E}[N(m)] \lesssim \sum_{i=1}^{m} \theta(r_i) r_i$; if $\theta(\varepsilon) = o(1/\varepsilon)$, we have $\theta(r_m) r_m = o(1)$, so that $\mathbb{E}[N(m)] = o(m)$. Thus, combining all the above observations, it seems the disagreement coefficient provides a reasonable quantification of the behavior and performance of CAL.

Although in general CAL is known to sometimes be suboptimal among active learning algorithms (see Section 8.2), there is a sense in which CAL *is* optimal among a certain family of selective sampling algorithms: namely, those that produce a complete correctly labeled data set (in the realizable case) by using $\hat{h}_{m-1}(X_m)$ in place of the true label for any labels not requested by the algorithm. Formally, we say a selective sampling algorithm is *perfect* if, for every $\mathcal{P}_{XY}$ in the realizable case, letting $(\{q_m\}_{m=1}^{\infty}, \{\hat{h}_m\}_{m=0}^{\infty})$ be the sequences produced by the algorithm (as above), $\forall m \in \mathbb{N}$, $\hat{h}_{m-1}(X_m) \neq Y_m \Rightarrow q_m = 1$. This family of algorithms was studied by El-Yaniv and Wiener [2010] (though from a somewhat different perspective, arising from the related *selective classification* problem), and the connection to the label complexity of active learning was later explored by El-Yaniv and Wiener [2012]. We can think of CAL as a perfect selective sampling algorithm: namely, the algorithm specified by $(\{q_m\}_{m=1}^{\infty}, \{\hat{h}_m\}_{m=0}^{\infty})$ with each $\hat{h}_m \in V_m^{\star}$ and each $q_m = \mathbb{1}_{\mathrm{DIS}(V_{m-1}^{\star})}(X_m)$.

Since the motivation for CAL was that we should request only those labels that cannot already be inferred from information already available, it is perhaps not surprising that CAL turns out to be *optimal* among perfect selective sampling algorithms in terms of the number of label requests. Specifically, a proof of El-Yaniv and Wiener [2010] implies that, for any perfect selective sampling algorithm $\mathcal{A}$, in the realizable case, with probability 1, for every $m \in \mathbb{N}$, $\mathcal{A}$ requests at least as many labels among $\{Y_i\}_{i=1}^{m}$ as CAL does. Essentially, if $X_i \in DIS\left(V_{i-1}^{\star}\right)$, there is a classifier $h \in V_{i-1}^{\star}$ for which $h(X_i) \neq f^{\star}(X_i)$, and we can construct an alternative $\mathcal{P}_{XY}$ where h is the target function, and under which there is a nonzero probability of getting this same $X_1, ..., X_i$ sequence, and hence the same $Y_1, ..., Y_{i-1}$ (since $h \in V_{i-1}^{\star}$), but $Y_i = f^{\star}(X_i)$ is necessarily different from the alternative Y_i label (i.e., $h(X_i)$), so that q_i must be 1 to satisfy the definition of perfect selective sampling. Furthermore, combined with Theorem 5.2, this also implies that any perfect selective sampling algorithm requests an expected number of labels among $\{Y_i\}_{i=1}^{m}$ at least

$\theta(1/m)/2$.

5.2 The Noisy Case

It is easy to see that CAL, as stated above, is not suitable for use in noisy settings. In particular, even the best classifier in $\mathbb{C}$ (namely, $f^\star$) may make some mistakes when there is noise, so that requesting even a single label Y_m for a point X_m with $f^\star(X_m) \neq Y_m$ may immediately preclude the possibility of CAL returning a classifier $\hat{h}$ with error rate close to that of $f^\star$ (e.g., this would be the case for threshold classifiers). Therefore, to make CAL robust to noise, we clearly need to change the way the set Q is used to constrain the $\hat{h}$ function. In keeping with the motivation for CAL (never requesting a label that could not possibly change the function $\hat{h}$ we return in the end), this amounts to changing the way the set V is defined in the second (equivalent) formulation of the algorithm.

The approach we discuss here is rooted in the seminal work of Balcan, Beygelzimer, and Langford [2006] on the so-called A^2 algorithm; due to the extensiveness of the literature on this subject, we defer a thorough survey of the development of this approach and the analysis thereof to Section 5.3 below. The basic strategy is motivated by two objectives. First, the update to V should maintain the invariant that $f^\star \in V$. Second, subject to this constraint, the update should use the requested labels to remove from V any classifiers that obviously have $\mathrm{er}(h) > \mathrm{er}(f^\star)$. One way to approach these objectives is to use the empirical error rates on the requested labels Q; in particular, both of these objectives would be satisfied if we were to replace the update to V in CAL by the update $V \leftarrow \{h \in V : \mathrm{er}_Q(h) > \mathrm{er}_Q(f^\star)\}$, where Q is the set of (X_i, Y_i) labeled data points for which $i \leq m$ and Y_i was requested by the algorithm. This is not quite feasible, since we do not have access to $\mathrm{er}_Q(f^\star)$. However, we can recover roughly the same type of behavior by invoking Lemma 3.1 to relate $\mathrm{er}_Q(h) - \min_{g \in V} \mathrm{er}_Q(g)$ to $\mathrm{er}(h) - \mathrm{er}(f^\star)$. In particular, note that for $h, g \in V$, any $i \leq m$ for which the algorithm did not request the label Y_i has $X_i \notin \mathrm{DIS}(V)$ anyway, so that $h(X_i) = g(X_i)$; but this implies

$$|Q|\left(\mathrm{er}_Q(h) - \min_{g\in V}\mathrm{er}_Q(g)\right) = m\left(\mathrm{er}_m(h) - \min_{g\in V}\mathrm{er}_m(g)\right).$$

Thus, by Lemma 3.1, with probability $1-\gamma$, if $f^\star \in V$ still, then $|Q|\left(\mathrm{er}_Q(f^\star) - \min_{g\in V}\mathrm{er}_Q(g)\right) \le mU(m,\gamma)$; therefore, we can safely remove any h from V that has $|Q|\left(\mathrm{er}_Q(h) - \min_{g\in V}\mathrm{er}_Q(g)\right) > mU(m,\gamma)$ as it must have $\mathrm{er}_Q(h) > \mathrm{er}_Q(f^\star)$. Lemma 3.1 has the further implication that the set V of classifiers that survive this update has $\sup_{h\in V}\mathrm{er}(h) - \mathrm{er}(f^\star) \le 2U(m,\gamma)$, so that as long as the algorithm processes enough unlabeled data points before halting, we will have a guarantee on the error rate of the returned classifier. There are a few subtleties being glossed over here, not the least of which is that the function $U(m,\gamma)$ is $\mathcal{P}_{XY}$-dependent, and we address these issues in more detail in the discussion below.

For $\delta \in (0,1)$ and $m \in \mathbb{N}$, define

$$\delta_m = \delta/(\log_2(2m))^2.$$

The specific algorithm we study here (a variant of the A^2 strategy of Balcan, Beygelzimer, and Langford, 2006, 2009) is stated formally as follows.

Algorithm: $\mathbf{RobustCAL}_\delta(n)$
0. $m \leftarrow 0$, $i \leftarrow 1$, $Q_i \leftarrow \{\}$
1. While $|Q_i| < n$ and $m < 2^n$
2. $\quad m \leftarrow m+1$
3. $\quad$ If, $\forall y \in \mathcal{Y}, \exists h \in \mathbb{C}$ with $h(X_m) = y$ and
$\qquad\qquad \forall j < i, (\mathrm{er}_{Q_j}(h) - \mathrm{er}^*_j)|Q_j| \le U(2^j, \delta_{(2^j)})2^j$
4. $\qquad$ Request the label Y_m; let $Q_i \leftarrow Q_i \cup \{(X_m, Y_m)\}$
5. $\quad$ If $\log_2(m) \in \mathbb{N}$
6. $\qquad \mathrm{er}^*_i \leftarrow \min\Big\{\mathrm{er}_{Q_i}(h) : h \in \mathbb{C}$ and
$\qquad\qquad\qquad \forall j < i, (\mathrm{er}_{Q_j}(h) - \mathrm{er}^*_j)|Q_j| \le U(2^j, \delta_{(2^j)})2^j\Big\}$
$\qquad i \leftarrow i+1$; $Q_i \leftarrow Q_{i-1}$
7. Return any $\hat{h} \in \mathbb{C}$ s.t. $\forall j < i, (\mathrm{er}_{Q_j}(\hat{h}) - \mathrm{er}^*_j)|Q_j| \le U(2^j, \delta_{(2^j)})2^j$

As was possible for CAL, we can write an equivalent form of this algorithm that makes the set V of surviving candidate classifiers explicit, which clarifies the connection to the motivation above, and simplifies the discussion in the proof below. Specifically, the following algorithm behaves identically to that stated above.

Algorithm: $\mathbf{RobustCAL}_\delta(n)$
0. $m \leftarrow 0$, $Q \leftarrow \{\}$, $V \leftarrow \mathbb{C}$
1. While $\lvert Q\rvert < n$ and $m < 2^n$
2. $\quad m \leftarrow m+1$
3. $\quad$ If $X_m \in \mathrm{DIS}(V)$
4. $\quad\quad$ Request the label Y_m; let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$
5. $\quad$ If $\log_2(m) \in \mathbb{N}$
6. $\quad\quad V \leftarrow \left\{h \in V : \left(\mathrm{er}_Q(h) - \min_{g \in V} \mathrm{er}_Q(g)\right) \lvert Q\rvert \le U(m, \delta_m)m\right\}$
7. Return any $\hat{h} \in V$

Note that, by induction, the set V is nonempty in Step 7, so that $\hat{h}$ is guaranteed to exist; specifically, given that V is nonempty going into Step 6, the $h \in V$ with minimal $\mathrm{er}_Q(h)$ remains in V after the update.

A few details of this algorithm deviate slightly from the motivation. First, the confidence arguments to $U(m, \cdot)$ vary with m; this is done in a way that makes the total failure probability sum up to at most δ in the proof below. Second, we are updating the set V only every time we double the number m of unlabeled samples processed, rather than for every m; though this certainly has computational advantages, our main motivation for doing this is the technical reason that it provides slightly better logarithmic factors in the label complexity guarantee below: namely, we get a $\log\log(1/\varepsilon)$ factor instead of $\log(1/\varepsilon)$, due to being able to take $\delta_m = \delta/(\log_2(2m))^2$ rather than something like $\delta/(1+m)^2$.

As noted, the above algorithm has a direct dependence on certain $\mathcal{P}_{XY}$-dependent values via the $U(\cdot,\cdot)$ function: namely, ν, α, a, and the $\theta(\cdot)$ function. The last of these can be removed by replacing θ with its trivial upper bound $\theta(r_0) \le 1/r_0$ in the definition of U, which only increases the label complexity below by logarithmic factors. However, another very elegant solution that removes *all* direct dependence on $\mathcal{P}_{XY}$

is provided by replacing $U(m, \delta_m)$ with a data-dependent bound. In fact, any reasonably-tight bound on $(\mathrm{er}_Q(f^\star) - \min_{g \in V} \mathrm{er}_Q(g))|Q|$ used in place of $U(m, \delta_m)m$ in the above algorithm will still lead to interesting behavior, and it is known that data-dependent bounds of this type exist which (in this context) can be bounded by a value proportional to $U(m, \delta_m)m$. In particular, bounds of this type (having no direct dependence on $\mathcal{P}_{XY}$), have been developed, for instance by Koltchinskii [2006], based on data-dependent Rademacher complexities. These data-dependent quantities have been used in active learning algorithms such as RobustCAL, for instance by Hanneke [2011, 2012] and Koltchinskii [2010]; in particular, Hanneke and Yang [2012] prove that the label complexity bound below also holds (up to constant factors) for a variant of RobustCAL that makes use of a data-dependent bound in place of $U(m, \delta_m)$ (in addition to other minor changes), and thus has no direct dependence on $\mathcal{P}_{XY}$. The essential motivation, strategy, and proof are not changed much by the addition of these data-dependent quantities in place of $U(m, \delta_m)$, and as such we will not go into the details of their definitions and properties here, so as to focus on the essential aspects of this active learning strategy and the analysis thereof; the interested reader is referred to the literature cited above for these details.

As implied by the motivation preceding the algorithm, the analysis of RobustCAL proceeds quite analogously to the analysis of CAL. Once again, the focus is on bounding $\sup_{h \in V} \mathrm{er}(h) - \mathrm{er}(f^\star)$ as a function of the number m of unlabeled data points processed, making use of Lemma 3.1 and either (3.1) or (3.2) to identify a sufficient size of m to guarantee this is at most ε with high probability. The problem then reduces to identifying a size of the budget n sufficient to reach this value of m before the number of label requests reaches the budget. This, in turn, boils down to bounding the sequence of probabilities of requesting the label Y_m, which can then be related to the sequence of radius(V) values via the disagreement coefficient. Finally, we can bound radius(V) in terms of $\sup_{h \in V} \mathrm{er}(h) - \mathrm{er}(f^\star)$, either via Condition 2.3, or in some cases by a simple triangle inequality argument. Since we already established a bound on $\sup_{h \in V} \mathrm{er}(h) - \mathrm{er}(f^\star)$ in the first step, that suffices to establish a label complexity bound. Working out the details of this line of reasoning leads to the following theorem.

Theorem 5.4. For any $\delta \in (0,1)$, RobustCAL$_\delta$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$, for a and α as in Condition 2.3, $\forall \varepsilon \in (0,1)$,

$$\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \tag{5.8}$$
$$a^2\theta\left(a\varepsilon^\alpha\right)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\mathrm{Log}\left(\theta\left(a\varepsilon^\alpha\right)\right)+\mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\right)\mathrm{Log}(1/\varepsilon),$$

and furthermore,

$$\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \tag{5.9}$$
$$\theta\left(\nu+\varepsilon\right)\left(\frac{\nu^2}{\varepsilon^2}+\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\left(d\mathrm{Log}\left(\theta(\nu+\varepsilon)\right)+\mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right).$$

Proof. Fix $\varepsilon, \delta \in (0,1)$, and consider running RobustCAL$_\delta$ with budget argument $n \in \mathbb{N}$ greater than

$$c'' \min\begin{cases} a^2\theta\left(a\varepsilon^\alpha\right)\varepsilon^{2(\alpha-1)}\left(d\mathrm{Log}\left(\theta\left(a\varepsilon^\alpha\right)\right)+\mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\right)\mathrm{Log}(1/\varepsilon) \\ \theta(\nu+\varepsilon)\left(\frac{\nu^2}{\varepsilon^2}+\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\left(d\mathrm{Log}\left(\theta\left(\nu+\varepsilon\right)\right)+\mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right)\end{cases}$$

for an appropriate numerical constant c'' (indicated by the analysis below). Proceeding as in the proof of Theorem 5.1, let $M \subseteq \{0, \ldots, 2^n\}$ denote the set of values of m obtained during the execution. For each $m \in M$, let V_m and Q_m denote the values of V and Q, respectively, upon reaching Step 1 with that value of m.

Lemma 3.1 and a union bound imply that, on an event E_δ of probability at least $1-\sum_{i=1}^{\infty}\frac{\delta}{(1+i)^2} > 1-2\delta/3$, every $m \in \mathbb{N}$ with $\log_2(m) \in \mathbb{N}$ has

$$\mathrm{er}_m(f^\star) - \min_{g\in\mathbb{C}} \mathrm{er}_m(g) \leq U(m, \delta_m), \tag{5.10}$$

and $\forall h \in \mathbb{C}$,

$$\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \max\left\{2(\mathrm{er}_m(h) - \mathrm{er}_m(f^\star)), U(m, \delta_m)\right\}. \tag{5.11}$$

In particular, since (as noted above) every $m \in M \setminus \{0\}$ and $h, g \in V_{m-1}$ have $\left(\mathrm{er}_{Q_m}(h) - \mathrm{er}_{Q_m}(g)\right)|Q_m| = \left(\mathrm{er}_m(h) - \mathrm{er}_m(g)\right)m$, if $f^\star \in V_{m-1}$

for some $m \in M$ with $\log_2(m) \in \mathbb{N}$, then

$$\forall h \in V_{m-1}, (\mathrm{er}_{Q_m}(h) - \mathrm{er}_{Q_m}(f^\star)) |Q_m| = (\mathrm{er}_m(h) - \mathrm{er}_m(f^\star)) m. \tag{5.12}$$

In particular, combined with (5.10), this implies

$$\left(\mathrm{er}_{Q_m}(f^\star) - \min_{g \in V_{m-1}} \mathrm{er}_{Q_m}(g)\right) |Q_m| \le U(m, \delta_m) m,$$

so that $f^\star \in V_m$. By induction (and the fact that $f^\star \in \mathbb{C}$, and $V_m = V_{m-1}$ when $\log_2(m) \notin \mathbb{N}$), we have that on the event E_δ, $f^\star \in V_m$ for all $m \in M$.

Now let $i_\varepsilon = \lceil \log_2(2/\varepsilon) \rceil$, define $I = \{0, \ldots, i_\varepsilon\}$, and for $i \in I$ let $\varepsilon_i = 2^{-i}$. For any $x \in (1, \infty)$, denote $\lceil x \rceil_2 = 2^{\lceil \log_2(x) \rceil}$: that is, the smallest power of 2 at least as large as x. Let $m_0 = 0$, and for c' as in (3.1) and (3.2), for each $i \in I \setminus \{0\}$, define

$$m_i' = \min \begin{cases} 4c' a \varepsilon_i^{\alpha - 2} \left(d\mathrm{Log}\left(\theta\left(a \varepsilon_i^\alpha\right)\right) + \mathrm{Log}\left(\frac{4 \log_2(4c' a / \varepsilon_i)}{\delta}\right)\right) \\ 4c' \left(\frac{\nu + \varepsilon_i}{\varepsilon_i^2}\right) \left(d\mathrm{Log}\left(\theta(\nu + \varepsilon_i)\right) + \mathrm{Log}\left(\frac{4 \log_2(4c' / \varepsilon_i)}{\delta}\right)\right) \end{cases}$$

and $m_i = \lceil m_i' \rceil_2$. In particular, for every $i \in I \setminus \{0\}$ with $m_i \in M$, combining (5.11), (5.12), the fact that $f^\star \in V_{m_i - 1}$, and the condition defining V in Step 6, we have that on the event E_δ,

$$\forall h \in V_{m_i}, \mathrm{er}(h) - \mathrm{er}(f^\star) \le 2U(m_i, \delta_{m_i}).$$

One can easily check that m_i satisfies the conditions of (3.1) and (3.2) with $\gamma = \delta_{m_i}$, so that $U(m_i, \delta_{m_i}) \le \varepsilon_i$ for all $i \in I \setminus \{0\}$. Combined with the fact that $er(h) - \mathrm{er}(f^\star) \le 2\varepsilon_0$ is trivially satisfied for every $h \in V_{m_0} = \mathbb{C}$, we have that on the event E_δ, every $i \in I$ with $m_i \in M$ satisfies

$$\forall h \in V_{m_i}, \mathrm{er}(h) - \mathrm{er}(f^\star) \le 2\varepsilon_i. \tag{5.13}$$

In particular, this implies that, to complete the proof, it suffices to show that $m_{i_\varepsilon} \in M$.

Next, turning to the analysis of the number of label requests, we can express the number of labels requested while $m \le m_{i_\varepsilon}$ as

$$\sum_{m=1}^{\min\{m_{i_\varepsilon}, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}(X_m) = \sum_{i=1}^{i_\varepsilon} \sum_{m=m_{i-1}+1}^{\min\{m_i, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}(X_m).$$

Now note that, by (5.13), on the event E_δ, for $i \in I \setminus \{0\}$ and $m \in \{m_{i-1}+1, \ldots, m_i\} \cap M$, $\mathrm{DIS}(V_{m-1}) \subseteq \mathrm{DIS}(V_{m_{i-1}}) \subseteq \mathrm{DIS}(\mathbb{C}(2\varepsilon_{i-1}))$, so that the above summation is at most

$$\sum_{i=1}^{i_\varepsilon} \sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}(\mathbb{C}(2\varepsilon_{i-1}))}(X_m).$$

This is a sum of independent Bernoulli random variables, so that a Chernoff bound implies that, on an event E'_δ of probability at least $1-\delta/3$, the value of the sum is at most

$$\log_2(3/\delta) + 2e \sum_{i=1}^{i_\varepsilon} (m_i - m_{i-1}) \mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(2\varepsilon_{i-1}\right)\right)\right). \tag{5.14}$$

Condition 2.3 implies that for $i \in I \setminus \{0\}$, $\mathbb{C}(2\varepsilon_{i-1}) \subseteq \mathrm{B}\left(f^\star, a(2\varepsilon_{i-1})^\alpha\right)$, so that $\mathcal{P}(\mathrm{DIS}(\mathbb{C}(2\varepsilon_{i-1}))) \leq \theta\left(a(2\varepsilon_{i-1})^\alpha\right) a(2\varepsilon_{i-1})^\alpha$. Furthermore, by a triangle inequality, we also have $\mathbb{C}(2\varepsilon_{i-1}) \subseteq \mathrm{B}\left(f^\star, 2\nu + 2\varepsilon_{i-1}\right)$, so that $\mathcal{P}(\mathrm{DIS}(\mathbb{C}(2\varepsilon_{i-1}))) \leq \theta\left(2(\nu+\varepsilon_{i-1})\right) 2(\nu+\varepsilon_{i-1})$. Combined with the definition of m_i, we have that for every $i \in I \setminus \{0\}$,

$$m_i \mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(2\varepsilon_{i-1}\right)\right)\right) \lesssim \tag{5.15}$$
$$\min \begin{cases} a^2 \theta\left(a(2\varepsilon_{i-1})^\alpha\right) \varepsilon_i^{2(\alpha-1)} \left(d \mathrm{Log}\left(\theta\left(a\varepsilon_i^\alpha\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon_i)}{\delta}\right)\right) \\ \theta\left(2(\nu+\varepsilon_{i-1})\right) \left(\frac{\nu^2}{\varepsilon_i^2} + 1\right) \left(d \mathrm{Log}\left(\theta\left(\nu+\varepsilon_i\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon_i)}{\delta}\right)\right) \end{cases}.$$

Plugging this into (5.14), and using basic properties of $\theta(\cdot)$ (namely, Theorem 7.1 and Corollary 7.2 from Chapter 7 below), combined with the fact that every $i \in I$ has $\varepsilon_i > \varepsilon/4$, we find that (5.14) is

$$\lesssim \min \begin{cases} a^2 \theta\left(a\varepsilon^\alpha\right) \varepsilon^{2(\alpha-1)} \left(d \mathrm{Log}\left(\theta\left(a\varepsilon^\alpha\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\right) \mathrm{Log}(1/\varepsilon) \\ \theta(\nu+\varepsilon) \left(\frac{\nu^2}{\varepsilon^2} + \mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right) \left(d \mathrm{Log}\left(\theta\left(\nu+\varepsilon\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right) \end{cases}.$$

For an appropriate choice of the constant c'', the budget n is larger than this, and furthermore $m_{i_\varepsilon} < 2^n$. Thus, we have proven that, on $E_\delta \cap E'_\delta$, we have $m_{i_\varepsilon} \in M$, so that $\hat{h} \in V_{m_{i_\varepsilon}}$. In particular, by (5.13), this implies that on $E_\delta \cap E'_\delta$, $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \sup_{h \in V_{m_{i_\varepsilon}}} \mathrm{er}(h) - \mathrm{er}(f^\star) \leq \varepsilon$. Furthermore, by a union bound, the event $E_\delta \cap E'_\delta$ has probability at least $1-\delta$. Noting that the above sufficient size of n matches the form of the bounds (5.8) and (5.9) from the theorem statement completes the proof. □

Recalling that passive learning can only achieve label complexities that are $\Omega(\varepsilon^{\alpha-2})$ for some distributions, this represents an improvement in label complexity when θ is small and $\alpha > 0$. Furthermore, even when $\alpha = 0$, the second bound on Λ in Theorem 5.4 can still reflect improvements in label complexity when ν and θ are particularly small. Again, these bounds are particularly interesting when $\theta(\varepsilon) = O(1)$.

It is known that the logarithmic factors in the bounds of Theorem 5.4 can sometimes be improved; specifically, by a careful treatment of the summation that results from plugging (5.15) into (5.14), one can replace the last $\mathrm{Log}(1/\varepsilon)$ factor by a constant in (5.8) when either $\alpha < 1$ or $\theta(\varepsilon) = \Theta(\varepsilon^{-\kappa})$ for some $\kappa \in (0,1]$. Similarly, when $\theta(\varepsilon) = \Theta(\varepsilon^{-\kappa})$, the $\mathrm{Log}(1/\varepsilon)$ term in (5.9) can be replaced by a constant (though this is really only interesting when ν is very small). It is also sometimes possible to remove the $\mathrm{LogLog}(1/\varepsilon)$ factor in both of these bounds (see the historical notes below for more on this).

Logarithmic factors aside, these bounds are sometimes fairly tight characterizations of the label complexity of RobustCAL. In particular, let $\varepsilon > 0$ be such that there exists a classifier $f \in \mathbb{C}$ with $\mathrm{er}(f) = \nu + 2\varepsilon$. Then, to guarantee $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \le \varepsilon$, we must guarantee $f \notin V$. Consider that Lemma 3.1 implies that (on the event from the proof of Theorem 5.4), since $f \in \mathbb{C}(2\varepsilon)$, we have $\mathrm{er}_m(f) - \min_{g\in\mathbb{C}} \mathrm{er}_m(g) \le \max\{4\varepsilon, U(m,\delta_m)\}$, so that RobustCAL must reach a value of m *at least* large enough to make $U(m,\delta_m) \le 4\varepsilon$ before f will be eliminated from V. Together, (3.1) and (3.2) represent a fairly tight inversion of $U(m,\gamma)$, so that the smallest m for which $U(m,\delta_m) \le 4\varepsilon$ is at least a value proportional to

$$\min\begin{cases} a\varepsilon^{\alpha-2}\left(d\mathrm{Log}\left(\theta\left(a\varepsilon^{\alpha}\right)\right) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)/\delta)\right) \\ \left(\frac{\nu+\varepsilon}{\varepsilon^2}\right)\left(d\mathrm{Log}\left(\theta(\nu+\varepsilon)\right) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)/\delta)\right)\end{cases}.$$

Letting m_ε denote this value, we have that, assuming it does not halt first, the number of labels requested by RobustCAL while $m \le m_\varepsilon$ (and therefore $f \in V$) is at least $\sum_{m=1}^{m_\varepsilon} \mathbb{1}_{\mathrm{DIS}(\mathbb{C}(2\varepsilon))}(X_m)$, which has expected value $m_\varepsilon \mathcal{P}(\mathrm{DIS}(\mathbb{C}(2\varepsilon)))$. Note that this is quite close to the value of the largest term in the summation in (5.14), and that the bounds of Theorem 5.4 are within logarithmic factors of an upper bound on this largest term. Thus, when the relaxations leading to this upper bound are tight (namely, the steps applying Condition 2.3 and bounding probability of disagreement in terms of the disagreement coefficient), the derived upper bound will be tight up to logarithmic factors.

Sometimes-Tightness It is known that the RobustCAL algorithm is sometimes itself suboptimal (see Chapter 8). However, it is still interesting to ask whether Theorem 5.4 can generally be improved, given that we commit to express our label complexity bounds only in terms of these quantities (i.e., ε, δ, d, θ, a, and α for (5.8), or ε, δ, d, θ, and ν for (5.9)). Toward addressing this, Raginsky and Rakhlin [2011] prove that certain dependences in (5.8) cannot be improved when $\alpha = 1$. Specifically, they show that for any $\varepsilon, \delta \in (0, 1/2)$, $a \in (1, \infty)$ (bounded away from one), sufficiently large $d \in \mathbb{N}$, and $\tau \in [2, 1/(a\varepsilon)]$, there exists an instance space $\mathcal{X}$ and hypothesis class $\mathbb{C}$ with $\mathrm{vc}(\mathbb{C}) = d$ such that, for any label complexity Λ achieved by an active learning algorithm, there exists a distribution $\mathcal{P}_{XY}$ satisfying Condition 2.3 with that a and with $\alpha = 1$ (in fact, satisfying (2.2)), with $\theta(a\varepsilon) = \tau$, for which

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \geq ca^2 \left(\theta(a\varepsilon)\mathrm{Log}(1/\delta) + d\mathrm{Log}(\theta(a\varepsilon))\right),$$

where $c \in (0, \infty)$ is a universal constant. This matches Theorem 5.4 (up to log factors) for this case, in terms of the asymptotic dependence on ε, as well as the dependence on a and d, though the bound in Theorem 5.4 multiplies these dependences instead of adding them. However, by a slight extension of the proof of Raginsky and Rakhlin [2011] (essentially, replicating the construction leading to the first term d independent times), one can strengthen the $ca^2 d\mathrm{Log}(\theta(a\varepsilon))$ term in this lower bound to $ca^2 d\theta(a\varepsilon)$; furthermore, this term is a valid lower bound for each $a \in [1, \infty)$, and even for the realizable case (though the term $ca^2\theta(a\varepsilon)\mathrm{Log}(1/\delta)$ is not).

These results are interesting for at least two reasons. One reason is that it means we can be assured that the above upper bounds are fairly tight in the case $\alpha = 1$, given that we have chosen to express them only in terms of these quantities. The other reason is that the construction of Raginsky and Rakhlin [2011] (and the aforementioned extension thereof) can be embedded in a variety of commonly-used hypothesis classes, including (homogeneous) linear separators in $\mathbb{R}^{3d}$ and axis-aligned rectangles in $\mathbb{R}^{2d}$, both of which have VC dimension $\propto d$, so that the above result implies a minimax lower bound for these hypothesis classes; in particular, taking $\tau = 1/(a\varepsilon)$, the stronger version of this bound implies a lower bound $\propto ad/\varepsilon$ on the minimax label complexity for these spaces, which is equivalent (up to logarithmic factors) to the minimax label complexity of passive learning for these problems (via Theorem 3.4).

5.3 Brief Survey of the Agnostic Active Learning Literature

Algorithms based on similar principles to RobustCAL have been explored in the active learning literature for a number of years now. This basic strategy is rooted in the seminal work of Balcan, Beygelzimer, and Langford [2006, 2009], who define an algorithm known as A^2, for *Agnostic Active* (adopting the terminology of the *Agnostic PAC* model of Kearns, Schapire, and Sellie, 1994). As with RobustCAL, the A^2 algorithm maintains a set of surviving candidate classifiers V, and only requests the labels of points in $\mathrm{DIS}(V)$. The main difference is that the update to V in A^2 is triggered by a test for whether $\mathcal{P}(\mathrm{DIS}(V))$ has been reduced by a factor of 2 since the last update; in contrast, RobustCAL updates V every time the number of processed unlabeled samples is doubled. The original work of Balcan, Beygelzimer, and Langford [2006, 2009] includes an analysis of the label complexity of A^2 for the space of threshold classifiers (Example 1) in terms of ε, δ, and ν, and an analysis for the space of homogeneous linear separators in $\mathbb{R}^k$ (under a certain uniform distribution) in terms of ε, δ, and k, which holds when $\nu \lesssim \varepsilon/\sqrt{k}$ (see Section 7.4). These results were later generalized by Hanneke [2007b] to all hypothesis classes, with a label complexity analysis expressed in terms of ε, δ, d, ν, and $\theta(\nu+\varepsilon)$. That bound is essentially the same as the result of (5.9) (up to logarithmic factors), except with the leading $\theta(\nu+\varepsilon)$ replaced by $\theta(\nu+\varepsilon)^2$.

The A^2 algorithm was in fact motivated slightly differently from the above, based on applying concentration inequalities for a data set sampled from the conditional distribution of $(X, Y) \sim \mathcal{P}_{XY}$ given $X \in \mathrm{DIS}(V)$. The above motivation, based on the fact that concentration inequalities for $(\mathrm{er}_m(h) - \min_{g \in V} \mathrm{er}_m(g))m$ also hold for $(\mathrm{er}_Q(h) - \min_{g \in V} \mathrm{er}_Q(g))|Q|$, is largely due to the later work of Dasgupta, Hsu, and Monteleoni [2007] (though, when viewed from an appropriate perspective, the two motivations can be seen as two sides of the same reasoning; see a recent result of Hanneke and Yang, 2012, for more on this). Dasgupta, Hsu, and Monteleoni [2007] define a slightly different algorithm from A^2, and bound its label complexity as a function of ε, δ, ν, and $\theta(\nu + \varepsilon)$; their bound essentially matches that of (5.9), except with a $\log^2(1/\varepsilon)$ where the above bound has a $\log(1/\varepsilon) \log\log(1/\varepsilon)$ factor; this small improvement in (5.9) is due to a combination of using a slightly better generalization bound $U(m, \delta_m)$, along with the aforementioned benefit from only updating every time $\log_2(m) \in \mathbb{N}$, and otherwise the result of Dasgupta, Hsu, and Monteleoni [2007] would be essentially identical to (5.9). In particular, the label complexity bound of Dasgupta, Hsu, and Monteleoni [2007] reduced the factor $\theta(\nu+\varepsilon)^2$ from the analysis of Hanneke [2007b] down to $\theta(\nu + \varepsilon)$.

These results all described their dependences on $\mathcal{P}_{XY}$ only via ν and $\theta(\cdot)$, and as such (in light of Theorem 4.3), necessarily have asymptotic dependence on ε of $\Omega(1/\varepsilon^2)$ when $\nu > 0$, and can reflect at best a factor of ν improvement in minimax label complexity compared to passive learning. Toward describing scenarios with a more interesting asymptotic improvement over passive learning, Castro and Nowak [2006, 2008] initiated the analysis of active learning under Condition 2.3. They specifically studied the problem of learning a threshold classifier (Example 1) under a special case of Condition 2.3, and found that label complexities on the order of $O\left(\varepsilon^{2(\alpha-1)} \vee \mathrm{Log}(1/\varepsilon)\right)$ are achievable for this problem, which matches the dependence in (5.8) for RobustCAL (with slightly better logarithmic factors, though see below about this); they also studied a certain nonparametric hypothesis class, known as boundary fragments (see Section 8.8).

The result of Castro and Nowak [2006, 2008] for thresholds was quickly extended by Balcan, Broder, and Zhang [2007] to the general problem of learning a (homogeneous) linear separator in $\mathbb{R}^k$ under a uniform distribution in a ball, under Condition 2.3. Balcan, Broder, and Zhang [2007] develop an algorithm specialized to this problem, and show that it achieves label complexities on the order of $a^2\varepsilon^{2(\alpha-1)}(d + \mathrm{Log}(1/\delta))\mathrm{polylog}(a/\varepsilon)$ for sufficiently small ε. Interestingly, this matches (up to logarithmic factors) the asymptotic dependence on ε in (5.8), but is smaller by the factor $\theta\,(a\varepsilon^{\alpha})$, which in this case is roughly $\propto \sqrt{d}$ for small ε (see Chapter 7), so that their algorithm embodies an interesting direction toward potentially improving the general theory of active learning under these conditions. Balcan, Broder, and Zhang [2007] additionally include an interesting analysis of the label complexity in infinite-dimensional spaces, under certain constraints on the distribution $\mathcal{P}$.

Castro and Langford subsequently claimed that the A^2 algorithm can be shown to achieve label complexities on the order of $O\left(\varepsilon^{2(\alpha-1)}\mathrm{polylog}(1/\varepsilon)\right)$ for the class of threshold classifiers under a special case of Condition 2.3 (personal communication), raising the question of whether a general analysis might be possible. The analysis of disagreement-based active learning under Condition 2.3 in the *general* case was initiated by Hanneke [2009a, 2011]. Specifically, that work analyzes the label complexity of both the A^2 algorithm of Balcan, Beygelzimer, and Langford [2006, 2009], and the method of Dasgupta, Hsu, and Monteleoni [2007] (with a modification to a bound used in the algorithm, which plays a role analogous to $U(m, \delta_m)$ in RobustCAL). That work finds that the original A^2 algorithm achieves a label complexity essentially identical to (5.8), except with $\theta\,(a\varepsilon^{\alpha})$ replaced by $\theta\,(a\varepsilon^{\alpha})^2$, and that the (modified) method of Dasgupta, Hsu, and Monteleoni [2007] achieves a label complexity essentially identical to (5.8) (up to logarithmic factors); as was the case for the original analysis of Dasgupta, Hsu, and Monteleoni [2007] compared to (5.9), the difference in logarithmic factors compared to (5.8) is essentially due to a combination of using a slightly different bound in place of $U(m, \delta_m)$, and updating on every round rather than only when $\log_2(m) \in \mathbb{N}$, and otherwise the label complexity of that method can be made to match (5.8) exactly. Hanneke [2011] additionally includes an analysis of active learning with classes of infinite VC dimension, the discussion of which we defer to Chapter 8.

The original analysis of the method of Dasgupta, Hsu, and Monteleoni [2007] by Hanneke [2009a, 2011] essentially contains all of the components of the above analysis of RobustCAL, though the process of applying those components to the algorithm of Dasgupta, Hsu, and Monteleoni [2007] makes that proof somewhat more involved than the proof of Theorem 5.4 above. It should also be noted that the aforementioned analysis of A^2 by Hanneke [2011] is an analysis of the original A^2 algorithm, as proposed by Balcan, Beygelzimer, and Langford [2006, 2009]. Interestingly, with essentially the same modifications to the bounds used in A^2 as Hanneke [2011] used in the analysis of the method of Dasgupta, Hsu, and Monteleoni [2007] (namely, data-dependent local Rademacher complexity bounds, which for our purposes, are at least as good as using $U(m, \delta_m)$), one can show that the A^2 algorithm does in fact achieve a label complexity satisfying (5.8).

Following up on the work of Hanneke [2009a, 2011], Koltchinskii [2010] proposed a related active learning algorithm (quite similar to RobustCAL, mainly differing in the condition that triggers an update to V). The analysis of Koltchinskii [2010] refines the work of Hanneke [2009a, 2011] in at least two respects. First, that method makes use of a simplified data-dependent threshold in the update of the set V (where $U(m, \delta_m)$ is used in RobustCAL). Second, the label complexity bound of Koltchinskii [2010] reflects an improvement by a logarithmic factor in the case $\alpha = 1$ compared to the result of Hanneke [2009a, 2011] (reducing a $\log(1/\varepsilon)$ factor to $\log\log(1/\varepsilon)$), again due to a combination of a tighter bound (or rather, analysis thereof) used to update V, along with the trick of only updating when $\log_2(m) \in \mathbb{N}$; in particular, the result of Koltchinskii [2010] for the case $\alpha = 1$ perfectly matches that of (5.8) for this case. In the process, Koltchinskii [2010] draws an interesting connection between disagreement-based active learning algorithms, such as RobustCAL, and the technique of *localization* from the literature on the label complexity of $\mathrm{ERM}(\mathbb{C}, \cdot)$ [e.g., Bartlett, Bousquet, and Mendelson, 2005, Koltchinskii, 2006]. This observation provides a perspective from which the analysis of disagreement-based active learning algorithms becomes an almost-mechanical process: bound the size m_i of m sufficient to guarantee $V \subseteq \mathbb{C}(2^{-i})$ given that $V \subseteq \mathbb{C}(2^{1-i})$ already, then sum $m_i \mathcal{P}(\mathrm{DIS}(\mathbb{C}(2^{1-i})))$ over $i \leq \lceil \log_2(1/\varepsilon) \rceil$. In particular, the style of proof presented above for Theorem 5.4 follows precisely this structure.

The exact RobustCAL algorithm stated above is essentially taken from Hanneke [2012] (though that work uses data-dependent bounds in place of $U(m, \delta_m)$, and has a slightly looser label complexity analysis). The label complexity analysis of RobustCAL given above is taken from the recent work of Hanneke and Yang [2012]. In fact, Hanneke and Yang [2012] also show that a slightly better result than Theorem 5.4 is achievable, which eliminates the $\log(1/\varepsilon)$ factor in the case $\alpha < 1$ (simply by a more careful treatment of the summations in the proof). They also find it is possible to eliminate the $\log\log(1/\varepsilon)$ factor in this case, by a careful choice of the δ_m confidence arguments, though the δ_m values achieving this re-introduce a dependence on $\mathcal{P}_{XY}$; it is presently unknown whether this improvement is generally achievable without any direct dependence on $\mathcal{P}_{XY}$.

There is another branch of the literature, also rooted in disagreement-based active learning, which seeks to make such algorithms more practically feasible. We discuss some of this work in more depth in Section 8.1, but for now, let us briefly survey the main idea. This line of work was initiated by Dasgupta, Hsu, and Monteleoni [2007], and furthered by the efforts of Beygelzimer, Hsu, Langford, and Zhang [2010] and Hsu [2010] [see also Beygelzimer, Hsu, Karampatziakis, Langford, and Zhang, 2011]. The intention is to avoid maintaining the set V, even in the form of constraints on the empirical error rates, since this introduces a computational overhead. The thinking is that, if we are careful not to introduce too much bias in the sample Q, the $h \in \mathbb{C}$ minimizing $\mathrm{er}_Q(h)$ should already be guaranteed to have small $\mathrm{er}(h) - \mathrm{er}(f^\star)$, so that the constraints on h would essentially be redundant anyway. The computational burden is then entirely on minimizing $\mathrm{er}_Q(h)$ (subject to

$h(X_m) = y$, in the equivalent of RobustCAL's Step 3), for which there are known heuristics from the passive learning literature. The algorithm of Beygelzimer, Hsu, Langford, and Zhang [2010] maintains this property of Q via an elegant importance-weighting trick introduced by Beygelzimer, Dasgupta, and Langford [2009], and then decides whether to request the label Y_m based on whether $\mathrm{er}_Q(h)$ has similar minimum values under each of the constraints $h(X_m) = +1$ and $h(X_m) = -1$; in essence, this is asking whether X_m is contained in a certain region of disagreement. This line of research has, so far, not produced label complexity bounds on the order of (5.8). However, the reasoning seems quite compatible with the analysis of RobustCAL, and it seems likely that it will eventually yield label complexities of similar magnitudes. In Chapter 6, we will discuss a modification of RobustCAL designed to address the computational challenge of optimizing and constraining the empirical error rate. This technique should be considered complimentary to the work of Beygelzimer, Hsu, Langford, and Zhang [2010], and the "final cut" of practically-useful disagreement-based active learning may likely take the form of a combination of these two ideas.

6

Computational Efficiency via Surrogate Losses

The previous sections have been almost exclusively focused on analyzing the *label complexity* of certain active learning methods. However, it is also quite important to have methods with reasonable *computational complexity*. An active learning method with good label complexity is still only useful in practice if its execution will terminate within a reasonable amount of time. However, several of the steps in Robust-CAL may often be computationally intractable to perform. For instance, even minimizing $\mathrm{er}_{Q_i}(h)$ over classifiers $h \in \mathbb{C}$ may be NP-Hard under certain noise conditions [see e.g., Guruswami and Raghavendra, 2009, Feldman, Gopalan, Khot, and Ponnuswami, 2009]. There is a developing literature on approaches to noise-robust computationally efficient (passive) learning, which attempts to achieve low computational complexity while maintaining a reasonable worst-case label complexity [e.g., Kearns, Schapire, and Sellie, 1994, Kalai, Klivans, Mansour, and Servedio, 2005, Feldman, Gopalan, Khot, and Ponnuswami, 2009]. However, relatively few results of this type have been obtained so far, and their applicability remains somewhat limited.

In the mean time, the applications community has embraced a *heuristic* approach to dealing with this computational hardness: namely, the use of convex *surrogate losses*. The reasoning is the following. If we let $\ell_{01}(z) = \mathbb{1}_{[-\infty,0]}(z)$ for $z \in [-\infty,\infty]$ (called the 0-1 loss), then for any $m \in \mathbb{N}$, $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^m$, and classifier h, we can express $\mathrm{er}_{\mathcal{L}}(h) = \frac{1}{m}\sum_{(x,y)\in\mathcal{L}} \ell_{01}(h(x)y)$. In this view, the difficulty of minimizing $\mathrm{er}_{\mathcal{L}}(h)$ over $h \in \mathbb{C}$ stems from the nonconvexity of ℓ_{01} and $\mathbb{C}$. To get around this problem, we can suppose every $h \in \mathbb{C}$ can be represented as $h = \mathrm{sign}(f)$, where $\mathrm{sign}(\cdot) = \mathbb{1}^{\pm}_{[0,\infty]}(\cdot)$, $f \in \mathcal{F}$, and $\mathcal{F}$ is some family of functions mapping $\mathcal{X}$ to $\mathbb{R}$. We might then replace $\ell_{01}(h(x)y)$ in the above average with the quantity $\ell(f(x)y)$, where ℓ is a well-chosen *convex* function. If this family $\mathcal{F}$ of functions is convex, then the average $\frac{1}{m}\sum_{(x,y)\in\mathcal{L}} \ell(f(x)y)$ typically *can* be efficiently optimized over the choice of $f \in \mathcal{F}$. In this context, the function ℓ is referred to as a *surrogate loss*.

This heuristic is at the core of almost every machine learning algorithm used in practice today; some methods are explicitly expressed as optimization problems with ℓ appearing in their objective functions (e.g., SVM), while others only implicitly optimize a surrogate loss via iterative descent (e.g., AdaBoost). But the general sense in practice today is that the choice of surrogate loss ℓ is as fundamental a part of the design of effective learning algorithms as the choice of hypothesis class or learning bias.

Though this heuristic has met with overwhelming success in most applications, it is clearly not always guaranteed to work, and often leads to methods that are not consistent (i.e., infinite label complexity) on distributions where the analogous (computationally-intractable) methods that directly optimize $\mathrm{er}_{\mathcal{L}}(h)$ would have quite reasonable label complexities. However, it is still possible to analyze the label complexities of these heuristic methods, under conditions that provably imply that the heuristic *will* work. Specifically, following Bartlett, Jordan, and McAuliffe [2006] and Zhang [2004], we might suppose the function $f^{\star}_{\ell} = \mathrm{argmin}_{f:\mathcal{X}\to[-\infty,\infty]} \mathbb{E}[\ell(f(X)Y)]$ is contained in $\mathcal{F}$, and furthermore satisfies $\mathrm{er}(\mathrm{sign}(f^{\star}_{\ell})) = \inf_{h:\mathcal{X}\to\mathcal{Y}} \mathrm{er}(h)$. It turns out this last condition is *always* satisfied, as long as the surrogate loss ℓ satisfies a simple condition (called classification calibration) described below. However, the condition that $f^\star_\ell \in \mathcal{F}$ is a much stronger requirement, and amounts to a constraint on the allowed $\mathcal{P}_{XY}$ distributions, mostly involving the form of the $\eta(\cdot)$ function.

In this chapter, we review the known results on the label complexities achievable by the use of surrogate losses in passive learning, and then continue by describing the analogous results obtained by Hanneke and Yang [2012] for a variant of RobustCAL that replaces ℓ_{01} with a surrogate loss ℓ in the various optimization steps, with appropriate modifications to the $U(m, \delta)$ function to compensate for this change. In particular, these results achieve the same *type* of improvement over the analogous passive learning method as was found for RobustCAL: multiplying the label complexity by a factor $\approx \theta(a\varepsilon^\alpha) \cdot a\varepsilon^\alpha$ under Condition 2.3.

6.1 Definitions and Notation

We will need a few more definitions to state and prove the results below. Specifically, we use the following notation. Let $\mathcal{F}$ be a set of measurable functions $f : \mathcal{X} \to \mathbb{R}$, called the *function class*. We will then suppose the hypothesis class $\mathbb{C}$ satisfies $\mathbb{C} = \{\mathrm{sign}(f) : f \in \mathcal{F}\}$. Though not technically necessary, it will simplify the discussion below to assume that there is a *bounded* measurable set $\bar{\mathcal{Y}} \subset \mathbb{R}$ such that every $f \in \mathcal{F}$ and $x \in \mathcal{X}$ satisfy $f(x) \in \bar{\mathcal{Y}}$ (see Hanneke and Yang, 2012, for the more general case); for convenience, we also suppose every $y \in \bar{\mathcal{Y}}$ has $-y \in \bar{\mathcal{Y}}$ as well, and to focus on nontrivial cases, we suppose $|\bar{\mathcal{Y}}| \geq 2$.

Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, and let $\ell : \bar{\mathbb{R}} \to [0, \infty]$ be any measurable function, called the *surrogate loss*; for convenience, we suppose $z \in \mathbb{R} \Rightarrow \ell(z) < \infty$, and that the quantity $\bar{\ell} = \sup_{z \in \bar{\mathcal{Y}}} \ell(z) \vee 1$ is bounded by some finite numerical constant. These assumptions are satisfied for most surrogate losses of interest for learning. To simplify the analysis here, we will not explicitly describe the dependence of the label complexity on the value $\bar{\ell}$ below (instead treating it as a numerical constant); the interested reader can find an explicit description of this dependence in the work of Hanneke and Yang [2012].

We use the following generalization of the notion of VC dimension to classes of real-valued functions. For any set $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$, let $\mathcal{G}_{\mathcal{H}} = \{\mathbb{1}^{\pm}_{\{((x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}: z<\ell(f(x)y))\}} : f \in \mathcal{H}\}$ denote the set of classifiers on $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ corresponding to the *subgraphs* of the functions $(x, y) \mapsto \ell(f(x)y)$ for functions $f \in \mathcal{H}$. Then define $d_\ell = \mathrm{vc}(\mathcal{G}_{\mathcal{F}})$, the VC dimension of $\mathcal{G}_{\mathcal{F}}$ (where, in this case, we consider $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ to be the instance space in the definition of the VC dimension); d_ℓ is called the *pseudo-dimension* [Pollard, 1990, Haussler, 1992]. For instance, if $\mathcal{F}$ is the class of linear functions $x \mapsto b + \sum_{i=1}^{k} w_i x_i$ defined over $x \in [-1, 1]^k$, where $b, w_1, \ldots, w_k \in [-r, r]$ for some $r \in (0, \infty)$, and ℓ is nonincreasing and nonconstant, then $d_\ell = k + 1$ [Dudley, 1987, Haussler, 1992]; furthermore, in the special case of $\mathcal{F} = \mathbb{C}$ and $\ell = \ell_{01}$, we have $d_\ell = d = \mathrm{vc}(\mathbb{C})$, the VC dimension of $\mathbb{C}$. Although Bartlett, Jordan, and McAuliffe [2006] and Hanneke and Yang [2012] explore a variety of combinations of function classes and losses, including some with $d_\ell = \infty$, for simplicity we restrict the present discussion to those scenarios with $d_\ell < \infty$; we discuss some results for scenarios with $d_\ell = \infty$ in Section 8.8; the interested reader is referred to the original works of Bartlett, Jordan, and McAuliffe [2006] and Hanneke and Yang [2012] for discussion of other general scenarios. For any set $\mathcal{H}$ of measurable functions $\mathcal{X} \to \mathbb{R}$, we also generalize the notion of the region of disagreement by defining $\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists f, g \in \mathcal{H} \text{ s.t. } \mathrm{sign}(f(x)) \neq \mathrm{sign}(g(x))\}$, the region of *sign* disagreement.

For any measurable function $f : \mathcal{X} \to \bar{\mathbb{R}}$ and probability measure P over $\mathcal{X} \times \mathcal{Y}$, define the *$\ell$-risk* $\mathrm{R}_\ell(f; P) = \mathbb{E}[\ell(f(X)Y)]$, where $(X, Y) \sim P$; when $P = \mathcal{P}_{XY}$, we abbreviate this as $\mathrm{R}_\ell(f) = \mathrm{R}_\ell(f; \mathcal{P}_{XY})$. For convenience, also overload the notation for error rate, defining $\mathrm{er}(f) = \mathcal{P}_{XY}((x, y) : \mathrm{sign}(f(x)) \neq y) = \mathrm{er}(\mathrm{sign}(f))$. Additionally, for $m \in \mathbb{N}$ and $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^m$, define $\mathrm{R}_\ell(f; \mathcal{L}) = \frac{1}{m} \sum_{(x,y)\in\mathcal{L}} \ell(f(x)y)$, the *empirical ℓ-risk*; for completeness, define $\mathrm{R}_\ell(f; \emptyset) = 0$. Furthermore, for any $\eta_0 \in [0, 1]$, define $\ell^\star(\eta_0) = \inf_{z\in\bar{\mathbb{R}}}(\eta_0\ell(z) + (1 - \eta_0)\ell(-z))$, and $\ell^\star_-(\eta_0) = \inf_{z\in\bar{\mathbb{R}}: z(2\eta_0-1)\leq 0}(\eta_0\ell(z) + (1 - \eta_0)\ell(-z))$. Note that $\ell^\star(\eta(x))$ represents the smallest possible value of $\mathbb{E}[\ell(zY)|X = x]$ over $z \in \bar{\mathbb{R}}$, where $(X, Y) \sim \mathcal{P}_{XY}$, while $\ell^\star_-(\eta(x))$ essentially represents the smallest such value, under the constraint that $\mathrm{sign}(z) \neq$

$\mathrm{sign}(\eta(x) - 1/2)$ (the Bayes optimal classification). In particular, this means $\inf_{f:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{R}_\ell(f) = \mathbb{E}[\ell^\star(\eta(X))]$. Though not strictly necessary for the main results below, for convenience we will suppose that for every $\eta_0 \in [0,1]$, the value $\ell^\star(\eta_0)$ is actually *attained* by some $z^\star(\eta_0) \in \bar{\mathbb{R}}$: that is, $\eta_0\ell(z^\star(\eta_0)) + (1-\eta_0)\ell(-z^\star(\eta_0)) = \ell^\star(\eta_0)$; this assumption will greatly simplify the discussion, and is always satisfied for most surrogate losses of interest anyway. We then define $f_\ell^\star(x) = z^\star(\eta(x))$ for every $x \in \mathcal{X}$. We therefore have

$$\begin{aligned}
\mathrm{R}_\ell(f_\ell^\star) = \mathbb{E}[\ell(z^\star(\eta(X))Y)] &= \mathbb{E}[\mathbb{E}[\ell(z^\star(\eta(X))Y)|X]] \\
&= \mathbb{E}[\eta(X)\ell(z^\star(\eta(X))) + (1-\eta(X))\ell(-z^\star(\eta(X)))] \\
&= \mathbb{E}[\ell^\star(\eta(X))] = \inf_{f:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{R}_\ell(f),
\end{aligned}$$

so that $f_\ell^\star$ is the global minimizer of R_ℓ. We will be particularly interested in surrogate losses ℓ for which any function f with $\mathrm{R}_\ell(f) = \mathrm{R}_\ell(f_\ell^\star)$ also minimizes the error rate $\mathrm{er}(f)$; in particular, this means $\eta(X) \neq 1/2 \Rightarrow \mathrm{sign}(f(X)) = \mathrm{sign}(\eta(X) - 1/2)$ with probability 1. A surrogate loss ℓ for which this is always true, regardless of $\mathcal{P}_{XY}$, is called *classification-calibrated*, following Bartlett, Jordan, and McAuliffe [2006]; this property can be equivalently characterized as follows.

Definition 6.1. ℓ is *classification-calibrated* if, $\forall \eta_0 \in [0,1] \setminus \{1/2\}$, $\ell_-^\star(\eta_0) > \ell^\star(\eta_0)$.

Bartlett, Jordan, and McAuliffe [2006] identify several interesting families of surrogate losses that are all classification-calibrated. For instance, they find that any convex loss with a strictly negative derivative at 0 is classification-calibrated (and in fact, that any convex loss without this property is not).

For any measurable functions f, g mapping $\mathcal{X}$ to $\mathbb{R}$, and any probability measure P over $\mathcal{X} \times \mathcal{Y}$, define $\mathrm{D}_\ell(f,g;P) = \sqrt{\mathbb{E}\left[(\ell(f(X)Y) - \ell(g(X)Y))^2\right]}$, where $(X,Y) \sim P$. The following condition on a distribution P over $\mathcal{X} \times \mathcal{Y}$ is essentially a natural generalization of Condition 2.3 (or rather, (2.1) and (2.2)) to general losses ℓ, and will be useful in stating our results below.

Condition 6.2. For $f^{\star}_{P,\ell} = \operatorname{argmin}_{g:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{R}_{\ell}(g;P)$, and some values $b \in [1,\infty)$ and $\beta \in [0,1]$, for every measurable function $f : \mathcal{X} \to \bar{\mathcal{Y}}$,

$$\mathrm{D}_{\ell}(f, f^{\star}_{P,\ell}; P)^2 \leq b \left(\mathrm{R}_{\ell}(f;P) - \mathrm{R}_{\ell}(f^{\star}_{P,\ell};P) \right)^{\beta}.$$

Any ℓ and P with $\mathrm{Var}(\ell(Y f^{\star}_{P,\ell}(X))) < \infty$ (where $(X,Y) \sim P$) have $\sup_{f:\mathcal{X}\to\bar{\mathcal{Y}}} \mathrm{D}_{\ell}(f, f^{\star}_{P,\ell}; P)^2 < \infty$, so that Condition 6.2 is trivially satisfied with $\beta = 0$ (interpreting $0^0 = 1$ in the context of Condition 6.2). Furthermore, it is easy to check that, in the case of $P = \mathcal{P}_{XY}$, $\ell = \ell_{01}$ (the 0-1 loss), and $\bar{\mathcal{Y}} = \mathcal{Y}$, Condition 6.2 is implied by (2.1), with $b = a$ and $\beta = \alpha$ in that case (and similarly for (2.2), with $\beta = 1$ in that case). However, for many scenarios, this condition has other interesting interpretations. In some cases, it does indeed place strong restrictions on the distribution P when $\beta > 0$. But for many commonly-used losses, Condition 6.2 is in fact *always* satisfied (under very mild noise conditions), and in these cases b and β merely represent quantities inherent in the definition of ℓ itself, independent of P. This fact is reflected in the following condition and lemma, due to Bartlett, Jordan, and McAuliffe [2006].

Condition 6.3. There exist constants $L \in [1,\infty)$, $C_{\ell} \in (0,\infty)$, and $r_{\ell} \in (0,\infty]$ such that $\forall x,y \in \bar{\mathcal{Y}}$, $|\ell(x)-\ell(y)| \leq L|x-y|$, and the function $\bar{\delta}_{\ell}(\varepsilon) = \inf\left\{ \frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) : x,y \in \bar{\mathcal{Y}}, |x-y| \geq \varepsilon \right\} \cup \{\infty\}$ satisfies $\forall \varepsilon \in [0,\infty)$, $\bar{\delta}_{\ell}(\varepsilon) \geq C_{\ell}\varepsilon^{r_{\ell}}$.

This condition essentially requires ℓ to be smooth and convex on the relevant range of possible arguments; the function $\bar{\delta}_{\ell}$ is referred to as the *modulus of convexity*. It is worth mentioning that all of the results concerning Condition 6.3 below continue to hold (with appropriate modification to constant factors) even if we replace the Euclidean metric (in the Lipschitz condition and definition of $\bar{\delta}_{\ell}$) with an arbitrary pseudometric bounded on $\bar{\mathcal{Y}}^2$, which thereby admits such losses as the truncated quadratic loss [Bartlett, Jordan, and McAuliffe, 2006]; indeed, one can show this generalization admits *every* ℓ that is convex and continuous (as well as the 0-1 loss, slightly modified so that $\ell_{01}(0) = 1/2$), though some necessarily have $r_{\ell} = \infty$. For simplicity, we will stick with this simpler condition (Euclidean metric) here, leaving the more general case as an exercise for the reader. Based on the above condition, we have the following lemma, which is a variant of a result proven by Bartlett, Jordan, and McAuliffe [2006].

Lemma 6.4. Suppose ℓ satisfies Condition 6.3. For any distribution P over $\mathcal{X} \times \mathcal{Y}$, letting $f^\star_{P,\ell} = \operatorname{argmin}_{g:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{R}_\ell(g;P)$, if $\{f^\star_{P,\ell}(x) : x \in \mathcal{X}\} \subseteq \bar{\mathcal{Y}}$, then P satisfies Condition 6.2 with values $\beta = \min\left\{1, \frac{2}{r_\ell}\right\}$ and $b = 2^{-\beta \min\{r_\ell - 1, 1\}} L^2 C_\ell^{-\beta} \left(\sup \bar{\mathcal{Y}}\right)^{-\beta \min\{r_\ell - 2, 0\}}$.

Proof. Let P be as in the lemma statement, fix any measurable function $f : \mathcal{X} \to \bar{\mathcal{Y}}$, and let $(X,Y) \sim P$. The result holds trivially if $r_\ell = \infty$, since then every $\varepsilon > 1$ has $\bar{\delta}_\ell(\varepsilon) = \infty$, which (together with the fact that $\sup_{y\in\bar{\mathcal{Y}}} \ell(y) < \infty$) implies $\sup_{x,y\in\bar{\mathcal{Y}}} |x-y| \le 1$, so that $\mathrm{D}_\ell(f, f^\star_{P,\ell}; P)^2 \le \sup_{x,y\in\bar{\mathcal{Y}}}(x-y)^2 \le 1 \le L^2 = b$. For the remainder of the proof, let us suppose $r_\ell < \infty$.

First note that

$$\begin{aligned}
&\frac{1}{2}\mathrm{R}_\ell(f;P) + \frac{1}{2}\mathrm{R}_\ell(f^\star_{P,\ell};P) - \mathrm{R}_\ell\left(\frac{1}{2}f + \frac{1}{2}f^\star_{P,\ell};P\right) \\
&= \mathbb{E}\left[\frac{1}{2}\ell(Yf(X)) + \frac{1}{2}\ell(Yf^\star_{P,\ell}(X)) - \ell\left(\frac{1}{2}Yf(X) + \frac{1}{2}Yf^\star_{P,\ell}(X)\right)\right] \\
&\ge \mathbb{E}\left[\bar{\delta}_\ell\left(\left|Yf(X) - Yf^\star_{P,\ell}(X)\right|\right)\right] = \mathbb{E}\left[\bar{\delta}_\ell\left(\left|f(X) - f^\star_{P,\ell}(X)\right|\right)\right] \\
&\qquad\qquad\qquad\qquad \ge C_\ell \mathbb{E}\left[\left|f(X) - f^\star_{P,\ell}(X)\right|^{r_\ell}\right].
\end{aligned}$$

When $r_\ell \ge 2$, Jensen's inequality implies

$$C_\ell \mathbb{E}\left[\left|f(X) - f^\star_{P,\ell}(X)\right|^{r_\ell}\right] \ge C_\ell \left(\mathbb{E}\left[\left(f(X) - f^\star_{P,\ell}(X)\right)^2\right]\right)^{r_\ell/2}.$$

On the other hand, when $r_\ell < 2$, $x \mapsto x^{r_\ell/2}$ is concave on $[0,\infty)$; thus, letting $\bar{B} = 2\sup\bar{\mathcal{Y}}$, for any $x \in [0, \bar{B}^2]$, Jensen's inequality implies $x^{r_\ell/2} = \left(\frac{x}{\bar{B}^2}\bar{B}^2 + \left(1 - \frac{x}{\bar{B}^2}\right)0\right)^{r_\ell/2} \ge \left(\frac{x}{\bar{B}^2}\right)\bar{B}^{r_\ell} + \left(1 - \frac{x}{\bar{B}^2}\right)0^{r_\ell/2} = \bar{B}^{r_\ell-2}x$. So in this case, $\left|f(X) - f^\star_{P,\ell}(X)\right|^{r_\ell} = \left(\left(f((X) - f^\star_{P,\ell}(X)\right)^2\right)^{r_\ell/2} \ge \bar{B}^{r_\ell-2}\left(f(X) - f^\star_{P,\ell}(X)\right)^2$, so that

$$C_\ell \mathbb{E}\left[\left|f(X) - f^\star_{P,\ell}(X)\right|^{r_\ell}\right] \ge C_\ell \bar{B}^{r_\ell-2} \mathbb{E}\left[\left(f(X) - f^\star_{P,\ell}(X)\right)^2\right].$$

In either case, we have established that

$$\frac{1}{2}\mathrm{R}_\ell(f;P) + \frac{1}{2}\mathrm{R}_\ell(f^\star_{P,\ell};P) - \mathrm{R}_\ell\left(\frac{1}{2}f + \frac{1}{2}f^\star_{P,\ell};P\right) \geq C_\ell \bar{B}^{\min\{r_\ell-2,0\}}\left(\mathbb{E}\left[\left(f(X) - f^\star_{P,\ell}(X)\right)^2\right]\right)^{1/\beta}. \quad (6.1)$$

By the Lipschitz guarantee in Condition 6.3,

$$\mathbb{E}\left[\left(f(X) - f^\star_{P,\ell}(X)\right)^2\right] = L^{-2}\mathbb{E}\left[\left(L\left|Yf(X) - Yf^\star_{P,\ell}(X)\right|\right)^2\right] \geq L^{-2}\mathbb{E}\left[\left(\left|\ell(Yf(X)) - \ell(Yf^\star_{P,\ell}(X))\right|\right)^2\right] = L^{-2}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^2.$$

Combined with (6.1), this implies

$$\frac{1}{2}\mathrm{R}_\ell(f;P) + \frac{1}{2}\mathrm{R}_\ell(f^\star_{P,\ell};P) - \mathrm{R}_\ell\left(\frac{1}{2}f + \frac{1}{2}f^\star_{P,\ell};P\right) \geq C_\ell \bar{B}^{\min\{r_\ell-2,0\}}L^{-2/\beta}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^{2/\beta}.$$

In particular, we have

$$\begin{aligned}
&\frac{1}{2}\mathrm{R}_\ell(f;P) + \frac{1}{2}\mathrm{R}_\ell(f^\star_{P,\ell};P) \\
&\geq \mathrm{R}_\ell\left(\frac{1}{2}f + \frac{1}{2}f^\star_{P,\ell};P\right) + C_\ell \bar{B}^{\min\{r_\ell-2,0\}}L^{-2/\beta}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^{2/\beta} \\
&\geq \mathrm{R}_\ell\left(f^\star_{P,\ell};P\right) + C_\ell \bar{B}^{\min\{r_\ell-2,0\}}L^{-2/\beta}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^{2/\beta},
\end{aligned}$$

where this last inequality follows by the minimality of $\mathrm{R}_\ell(f^\star_{P,\ell};P)$. Subtracting $\mathrm{R}_\ell(f^\star_{P,\ell};P)$ from the first and last expressions, and then multiplying by 2, we have

$$\mathrm{R}_\ell(f;P) - \mathrm{R}_\ell(f^\star_{P,\ell};P) \geq 2C_\ell \bar{B}^{\min\{r_\ell-2,0\}}L^{-2/\beta}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^{2/\beta} = \left(\frac{1}{b}\mathrm{D}_\ell(f, f^\star_{P,\ell};P)^2\right)^{1/\beta},$$

which implies Condition 6.2 with the given values of b and β. □

6.2 Bounding Excess Error Rate with Excess Surrogate Risk

Bartlett, Jordan, and McAuliffe [2006] study a general technique for converting excess ℓ-risk guarantees into guarantees on the excess error rate, when ℓ is classification-calibrated. Specifically, for $z \in [-1, 1]$, define $\tilde{\psi}_\ell(z) = \ell^\star_-\left(\frac{1+z}{2}\right) - \ell^\star\left(\frac{1+z}{2}\right)$, and let ψ_ℓ be the largest convex lower bound of $\tilde{\psi}_\ell$ on $[0,1]$; for convenience, also define $\psi_\ell(z)$ for $z \in (1,\infty)$ as any value that maintains the convexity of the function ψ_ℓ on $[0,\infty)$, and continuity on $[1,\infty)$. Bartlett, Jordan, and McAuliffe [2006] show that the function ψ_ℓ is continuous on $(0,1)$ and nondecreasing on $(0,\infty)$, and in fact that (since it is a nonnegative convex function with $\psi_\ell(0) = 0$) $z \mapsto \psi_\ell(z)/z$ is nondecreasing on $(0,\infty)$ as well; note that this also implies that, for any $c > 0$, $z \mapsto \psi_\ell(cz)/z$ is nondecreasing on $(0,\infty)$, since it is proportional to $\psi_\ell(cz)/(cz)$; equivalently, we have that $z \mapsto z\psi_\ell(c/z)$ is nonincreasing on $(0,\infty)$ for any $c > 0$. Bartlett, Jordan, and McAuliffe [2006] further show that, if ℓ is classification-calibrated, then ψ_ℓ is *strictly* increasing on $(0,\infty)$. Additionally, they prove that every measurable function $f : \mathcal{X} \to \bar{\mathbb{R}}$ satisfies $\psi_\ell(\mathrm{er}(f) - \mathrm{er}(f^\star_\ell)) \le \mathrm{R}_\ell(f) - \mathrm{R}_\ell(f^\star_\ell)$. For our purposes, we will make use of a stronger result for classification-calibrated losses, holding under Condition 2.3. Specifically, letting a and α be values satisfying Condition 2.3, $\forall \varepsilon > 0$ define

$$\Psi_\ell(\varepsilon) = a\varepsilon^\alpha \psi_\ell\left(\varepsilon^{1-\alpha}/(2a)\right).$$

Note that, since $z \mapsto \psi_\ell(z/(2a))/z$ is nondecreasing on $(0,\infty)$, for any $x, y \in (0,\infty)$ with $x < y$, $\Psi_\ell(x) = ax\left(\psi_\ell\left(x^{1-\alpha}/(2a)\right)/x^{1-\alpha}\right) \le ax\left(\psi_\ell\left(y^{1-\alpha}/(2a)\right)/y^{1-\alpha}\right) \le \Psi_\ell(y)$, so that Ψ_ℓ is nondecreasing on $(0,\infty)$ (and in fact, strictly increasing if ℓ is classification-calibrated, since the last inequality above is strict in that case). The function Ψ_ℓ will be used in the statements of most of our results in this section; as the following lemma [due to Bartlett, Jordan, and McAuliffe, 2006] indicates, it provides a way to convert guarantees on the ℓ-risk of a function into guarantees on the error rate of the corresponding classifier.

Lemma 6.5. Suppose ℓ is classification-calibrated, $f^\star_\ell \in \mathcal{F}$, and that $\mathcal{P}_{XY}$ satisfies Condition 2.3 for given a and α values. For any $\varepsilon > 0$ and any measurable function $f : \mathcal{X} \to \bar{\mathbb{R}}$ with $\mathrm{sign}(f) \in \mathbb{C}$,

$$\mathrm{R}_\ell(f) - \mathrm{R}_\ell(f^\star_\ell) \le \Psi_\ell(\varepsilon) \implies \mathrm{er}(f) - \mathrm{er}(f^\star) \le \varepsilon.$$

Proof. Let $f : \mathcal{X} \to \bar{\mathbb{R}}$ be a measurable function. If $\mathrm{er}(f) - \mathrm{er}(f^\star) = 0$, the result is trivially satisfied for all $\varepsilon > 0$; for the remainder of the proof, suppose $\mathrm{er}(f) - \mathrm{er}(f^\star) > 0$. Since ℓ is classification-calibrated, we have $\mathrm{sign}(f_\ell^\star(\cdot)) = \mathrm{sign}(\eta(\cdot) - 1/2)$, and therefore $\mathrm{er}(f_\ell^\star) = \inf_{g:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{er}(g)$; in particular, since $f_\ell^\star \in \mathcal{F}$, this implies $f^\star(\cdot) = \mathrm{sign}(f_\ell^\star(\cdot)) = \mathrm{sign}(\eta(\cdot) - 1/2)$ as well. Therefore, letting $(X,Y) \sim \mathcal{P}_{XY}$, for every $t \in (0,1]$ we have

$$\begin{aligned} \mathrm{er}(f) - \mathrm{er}(f^\star) &= \mathbb{E}\left[\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f), f^\star\})}(X)|2\eta(X) - 1|\right] \\ &\le t\mathcal{P}(x : \mathrm{sign}(f(x)) \neq f^\star(x)) \\ &\quad + \mathbb{E}\left[\mathbb{1}_{[t,1]}(|2\eta(X) - 1|)\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f), f^\star\})}(X)|2\eta(X) - 1|\right]. \end{aligned} \tag{6.2}$$

For $x \in \mathcal{X}$ with $|2\eta(x) - 1| \ge t$, we have $\psi_\ell(|2\eta(x) - 1|) > 0$ (since ψ_ℓ is strictly increasing for classification-calibrated ℓ), so that $|2\eta(x) - 1| = \frac{|2\eta(x)-1|}{\psi_\ell(|2\eta(x)-1|)}\psi_\ell(|2\eta(x) - 1|)$; then the facts that $z \mapsto \psi_\ell(z)/z$ is nondecreasing on $(0,1]$ and ψ_ℓ is nonnegative imply

$$\frac{|2\eta(x) - 1|}{\psi_\ell(|2\eta(x) - 1|)}\psi_\ell(|2\eta(x) - 1|) \le \frac{t}{\psi_\ell(t)}\psi_\ell(|2\eta(x) - 1|).$$

Since ψ_ℓ is nonnegative, we have that

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}_{[t,1]}(|2\eta(X) - 1|)\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f), f^\star\})}(X)|2\eta(X) - 1|\right] \\ &\qquad \le \mathbb{E}\left[\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f), f^\star\})}(X)\frac{t}{\psi_\ell(t)}\psi_\ell(|2\eta(X) - 1|)\right]. \end{aligned} \tag{6.3}$$

We generally have $\psi_\ell(|2\eta(x) - 1|) \le \tilde{\psi}_\ell(|2\eta(x) - 1|) = \ell_-^\star(\max\{\eta(x), 1 - \eta(x)\}) - \ell^\star(\max\{\eta(x), 1 - \eta(x)\})$. Noting that $\ell_-^\star(1 - \eta(x)) = \ell_-^\star(\eta(x))$ and $\ell^\star(1 - \eta(x)) = \ell^\star(\eta(x))$, we have $\psi_\ell(|2\eta(x) - 1|) \le \ell_-^\star(\eta(x)) - \ell^\star(\eta(x))$. If $\mathrm{sign}(f(x)) \neq f^\star(x)$, then we must have $f(x)(2\eta(x) - 1) \le 0$, so that (by definition of $\ell_-^\star$), $\ell_-^\star(\eta(x)) \le \eta(x)\ell(f(x)) + (1 - \eta(x))\ell(-f(x))$. Furthermore, recalling the definition of $f_\ell^\star$, we have $\eta(x)\ell(f_\ell^\star(x)) + (1 - \eta(x))\ell(-f_\ell^\star(x)) = \eta(x)\ell(z^\star(\eta(x))) + (1 - \eta(x))\ell(-z^\star(\eta(x))) = \ell^\star(\eta(x))$. Altogether, we have that when $\mathrm{sign}(f(x)) \neq f^\star(x)$, $\psi_\ell(|2\eta(x) - 1|) \le \eta(x)\left(\ell(f(x)) - \ell(f_\ell^\star(x))\right) + (1 - \eta(x))\left(\ell(-f(x)) - \ell(-f_\ell^\star(x))\right) = \mathbb{E}\left[\ell(Yf(X)) - \ell(Yf_\ell^\star(X))|X = x\right]$. Applying this to the expression above, we find that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f),f^\star\})}(X)\frac{t}{\psi_\ell(t)}\psi_\ell(|2\eta(X)-1|)\right]\\&\quad\leq\frac{t}{\psi_\ell(t)}\mathbb{E}\left[\mathbb{1}_{\mathrm{DIS}(\{\mathrm{sign}(f),f^\star\})}(X)\mathbb{E}\left[\left(\ell(Yf(X))-\ell(Yf^\star_\ell(X))\right)|X\right]\right].\end{aligned}$$

By definition of $f^\star_\ell$, $\mathbb{E}\left[\ell(Yf(X))-\ell(Yf^\star_\ell(X))|X\right]$ is nonnegative, so that the above expression is at most

$$\begin{aligned}&\frac{t}{\psi_\ell(t)}\mathbb{E}\left[\mathbb{E}\left[\ell(Yf(X))-\ell(Yf^\star_\ell(X))|X\right]\right]\\&\qquad=\frac{t}{\psi_\ell(t)}\mathbb{E}\left[\ell(Yf(X))-\ell(Yf^\star_\ell(X))\right]=\frac{t}{\psi_\ell(t)}\left(\mathrm{R}_\ell(f)-\mathrm{R}_\ell(f^\star_\ell)\right).\end{aligned}$$

Plugging this into (6.3), and combining with (6.2), we have that

$$\mathrm{er}(f)-\mathrm{er}(f^\star)\leq t\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))+\frac{t}{\psi_\ell(t)}\left(\mathrm{R}_\ell(f)-\mathrm{R}_\ell(f^\star_\ell)\right).$$

Noting that $\mathrm{er}(f)-\mathrm{er}(f^\star)\leq\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))$, and setting $t=\frac{\mathrm{er}(f)-\mathrm{er}(f^\star)}{2\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))}$, after a bit of algebra we find that

$$\begin{aligned}&\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))\psi_\ell\left(\frac{\mathrm{er}(f)-\mathrm{er}(f^\star)}{2\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))}\right)\\&\qquad\qquad\leq\mathrm{R}_\ell(f)-\mathrm{R}_\ell(f^\star_\ell).\end{aligned}\tag{6.4}$$

Since $\mathrm{sign}(f)\in\mathbb{C}$, Condition 2.3 implies $\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))\leq a(\mathrm{er}(f)-\mathrm{er}(f^\star))^\alpha$. Together with the fact that $z\mapsto z\psi_\ell(c/z)$ is nonincreasing on $(0,\infty)$ for any $c>0$, this implies

$$\begin{aligned}&a\left(\mathrm{er}(f)-\mathrm{er}(f^\star)\right)^\alpha\psi_\ell\left(\frac{\mathrm{er}(f)-\mathrm{er}(f^\star)}{2a\left(\mathrm{er}(f)-\mathrm{er}(f^\star)\right)^\alpha}\right)\\&\qquad\leq\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))\psi_\ell\left(\frac{\mathrm{er}(f)-\mathrm{er}(f^\star)}{2\mathcal{P}(x:\mathrm{sign}(f(x))\neq f^\star(x))}\right).\end{aligned}$$

Noting that the left side of this inequality equals $\Psi_\ell\left(\mathrm{er}(f)-\mathrm{er}(f^\star)\right)$, and combining with (6.4), we have

$$\Psi_\ell\left(\mathrm{er}(f)-\mathrm{er}(f^\star)\right)\leq\mathrm{R}_\ell(f)-\mathrm{R}_\ell(f^\star_\ell).$$

name	$\ell(z)$	$\Psi_\ell(\varepsilon)$	L	C_ℓ	r_ℓ
quadratic	$(1-z)^2$	$\frac{\varepsilon^{2-\alpha}}{4a}$	$2+2\sup\bar{\mathcal{Y}}$	$1/4$	2
exponential	e^{-z}	$\approx \frac{\varepsilon^{2-\alpha}}{8a}$	$e^{\sup\bar{\mathcal{Y}}}$	$\frac{e^{-\sup\bar{\mathcal{Y}}}}{8}$	2
hinge	$\max\{1-z,0\}$	$\varepsilon/2$	1	N/A	N/A
0-1	$\mathbb{1}_{[-\infty,0]}(z)$	$\varepsilon/2$	1	N/A	N/A

Table 6.1: This table lists several commonly-used loss functions, along with the associated quantities defined above.

Therefore, for any $\varepsilon > 0$ with $\mathrm{R}_\ell(f) - \mathrm{R}_\ell(f_\ell^\star) \le \Psi_\ell(\varepsilon)$, we also have $\Psi_\ell(\mathrm{er}(f) - \mathrm{er}(f^\star)) \le \Psi_\ell(\varepsilon)$; since ℓ being classification-calibrated implies Ψ_ℓ is strictly increasing on $(0,\infty)$, this implies $\mathrm{er}(f) - \mathrm{er}(f^\star) \le \varepsilon$. $\square$

6.3 Examples

Here we review a few examples of loss functions ℓ commonly used in machine learning. These examples are taken from the work of Bartlett, Jordan, and McAuliffe [2006]. The relevant quantities are summarized in Table 6.1.

For comparison, we first calculate the above quantities for the 0-1 loss itself. In this case, recall $\ell(z) = \ell_{01}(z) = \mathbb{1}_{[-\infty,0]}(z)$. Supposing $\bar{\mathcal{Y}} = \mathcal{Y}$, every $x, y \in \bar{\mathcal{Y}}$ have $|\ell(x) - \ell(y)| \in \{0,1\}$, with $|\ell(x)-\ell(y)| = 1$ only if $x = -y$, so that $|\ell(x) - \ell(y)| \le |x-y|/2$; thus, the 0-1 loss does satisfy the Lipschitz requirement in Condition 6.3, with $L = 1$. However, $\frac{1}{2}\ell(1) + \frac{1}{2}\ell(-1) - \ell(\frac{1}{2}(1) + \frac{1}{2}(-1)) = \frac{1}{2} - \ell(0) = -\frac{1}{2}$, so that the 0-1 loss does not satisfy the condition on $\bar{\delta}_\ell$ required by Condition 6.3. Since every $z \in \bar{\mathbb{R}}$ has either $\ell(z) = 1$ or $\ell(-z) = 1$ (and both are 1 only for $z = 0$), we have $\ell^\star(\eta_0) = \min\{\eta_0, 1-\eta_0\}$; furthermore, any z with $z(2\eta_0 - 1) < 0$ has $\ell(z) = \ell(-z^\star(\eta_0))$ and $\ell(-z) = \ell(z^\star(\eta_0))$; combined with the facts that $\ell_-^\star(1/2) = \ell^\star(1/2)$ and $\eta_0\ell(0) + (1-\eta_0)\ell(0) = 1 \ge \ell_-^\star(\eta_0)$, this implies $\ell_-^\star(\eta_0) = \max\{\eta_0, 1-\eta_0\}$. In particular, since $\max\{\eta_0, 1-\eta_0\} > \min\{\eta_0, 1-\eta_0\}$ for any $\eta_0 \in [0,1] \setminus \{1/2\}$, we see that ℓ_{01} is indeed classification-calibrated. Furthermore, for $z \in [0,1]$, $\tilde{\psi}_\ell(z) = \ell_-^\star\left(\frac{1+z}{2}\right) - \ell^\star\left(\frac{1+z}{2}\right) = \frac{1+z}{2} - \frac{1-z}{2} = z$; since this is already convex, we have $\psi_\ell(z) = z$ as well. Therefore, for any $\varepsilon \in [0,1]$, $\Psi_\ell(\varepsilon) = a\varepsilon^\alpha \psi_\ell\left(\varepsilon^{1-\alpha}/(2a)\right) = \varepsilon/2$.

Example 6.1. $\ell(z) = (1-z)^2$, the *quadratic loss.*
The quadratic loss is used in many popular machine learning methods, including the classic work on supervised training of multilayer neural networks by back-propagation of errors [Rumelhart, Hinton, and Williams, 1986]. This loss has a particularly appealing property for theoretical work: namely, that $f^\star_\ell(x) = 2\eta(x)-1$ for all $x \in \mathcal{X}$, which can be observed by differentiating $\eta_0\ell(z) + (1-\eta_0)\ell(-z) = \eta_0(1-z)^2 + (1-\eta_0)(1+z)^2$ with respect to z to arrive at $2(1-\eta_0)(1+z) - 2\eta_0(1-z) = 2(1+z) - 4\eta_0$, and then setting this equal to 0 and solving to find $z^\star(\eta_0) = 2\eta_0 - 1$. For instance, this fact makes it particularly easy to determine whether a given distribution (specified in terms of $\mathcal{P}$ and η) satisfies $f^\star_\ell \in \mathcal{F}$, a condition we will rely on in many of the results below. This also implies $\ell^\star(\eta_0) = \eta_0(2-2\eta_0)^2 + (1-\eta_0)(2\eta_0)^2 = 4\eta_0(1-\eta_0)$. The form of the derivative above further reveals that $\eta_0\ell(z) + (1-\eta_0)\ell(-z)$ is strictly increasing in $|z - z^\star(\eta_0)|$, so that among $z \in \bar{\mathbb{R}}$ with $z(2\eta_0 - 1) \le 0$, the value of z minimizing $\eta_0\ell(z) + (1-\eta_0)\ell(-z)$ is 0, which therefore implies $\ell^\star_-(\eta_0) = \eta_0\ell(0) + (1-\eta_0)\ell(0) = 1$. In particular, any $\eta_0 \in [0,1] \setminus \{1/2\}$ has $\ell^\star_-(\eta_0) = 1 > 4\eta_0(1-\eta_0) = \ell^\star(\eta_0)$, so that ℓ is classification-calibrated. Furthermore, we have that for $z \in [0,1]$, $\tilde{\psi}_\ell(z) = 1 - 4\left(\frac{1+z}{2}\right)\left(\frac{1-z}{2}\right) = z^2$. This is already a convex function, so we also have $\psi_\ell(z) = z^2$. Thus, for $\varepsilon \in [0,1]$, we have $\Psi_\ell(\varepsilon) = a\varepsilon^\alpha\left(\varepsilon^{1-\alpha}/(2a)\right)^2 = \frac{\varepsilon^{2-\alpha}}{4a}$.

Additionally, for any $z \in [\inf\bar{\mathcal{Y}}, \sup\bar{\mathcal{Y}}] = [-\sup\bar{\mathcal{Y}}, \sup\bar{\mathcal{Y}}]$, we can bound the magnitude of the derivative $|\ell'(z)| = 2|1-z| \le 2 + 2\sup\bar{\mathcal{Y}}$; this implies any $x, y \in \bar{\mathcal{Y}}$ satisfy $|\ell(x) - \ell(y)| \le \left(2 + 2\sup\bar{\mathcal{Y}}\right)|x-y|$, so that ℓ satisfies the Lipschitz requirement in Condition 6.3 with $L = 2 + 2\sup\bar{\mathcal{Y}}$. Furthermore, for any $x, y \in \bar{\mathcal{Y}}$, $\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) = \frac{1}{2}(1-x)^2 + \frac{1}{2}(1-y)^2 - (\frac{1}{2}(1-x) + \frac{1}{2}(1-y))^2 = \frac{1}{4}(1-x)^2 + \frac{1}{4}(1-y)^2 - \frac{1}{2}(1-x)(1-y) = \frac{1}{4}(y-x)^2$, which implies, $\forall\varepsilon \ge 0$, $\bar{\delta}_\ell(\varepsilon) \ge \frac{1}{4}\varepsilon^2$, so that ℓ satisfies the requirement on the modulus of convexity in Condition 6.3 with $C_\ell = 1/4$ and $r_\ell = 2$.

Example 6.2. $\ell(z) = e^{-z}$, the *exponential loss.*
The exponential loss plays a key role in many machine learning methods, such as AdaBoost. In this case, ℓ is differentiable, with derivative $\ell'(x) = -e^{-x}$, and $|\ell'(x)|$ is maximized over $x \in [\inf \bar{\mathcal{Y}}, \sup \bar{\mathcal{Y}}]$ at $x = \inf \bar{\mathcal{Y}}$ with value $|\ell'(\inf \bar{\mathcal{Y}})| = e^{-\inf \bar{\mathcal{Y}}} = e^{\sup \bar{\mathcal{Y}}}$. Therefore, any $x, y \in \bar{\mathcal{Y}}$ with $x \leq y$ have $\ell(x) \leq \ell(y) + e^{\sup \bar{\mathcal{Y}}}(y - x)$ so that $|\ell(x) - \ell(y)| \leq e^{\sup \bar{\mathcal{Y}}}|x - y|$, which means ℓ satisfies the Lipschitz requirement in Condition 6.3 with $L = e^{\sup \bar{\mathcal{Y}}}$. Furthermore, we have $\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) = \frac{1}{2}\left(e^{-x} + e^{-y}\right) - e^{-(x+y)/2} = \frac{1}{2}e^{-y}\left(e^{(y-x)} - 2e^{(y-x)/2} + 1\right) = \frac{1}{2}e^{-y}\left(e^{(y-x)/2} - 1\right)^2$. For any differentiable convex function g, with derivative g', and any $c, c' \in \mathbb{R}$, $g(c') - g(c) \geq g'(c)(c' - c)$. Applying this with $g(z) = e^z = g'(z)$, $c' = (y - x)/2$, and $c = 0$, we have $e^{(y-x)/2} - 1 = e^{(y-x)/2} - e^0 \geq e^0((y - x)/2 - 0) = (y - x)/2$. Since $y \geq x$ implies $(y - x)/2 \geq 0$, we have $\left(e^{(y-x)/2} - 1\right)^2 \geq (y - x)^2/4$. Plugging this into the above, we have $\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) \geq \frac{1}{8}e^{-y}(y - x)^2 \geq \frac{1}{8}e^{-\sup \bar{\mathcal{Y}}}(y - x)^2$. Thus, $\forall \varepsilon \geq 0$, $\bar{\delta}_\ell(\varepsilon) \geq \frac{1}{8}e^{-\sup \bar{\mathcal{Y}}}\varepsilon^2$, so that ℓ satisfies the requirement on $\bar{\delta}_\ell(\cdot)$ in Condition 6.3 with $C_\ell = \frac{1}{8}e^{-\sup \bar{\mathcal{Y}}}$ and $r_\ell = 2$.

Next, toward calculating $\ell^\star(\eta_0)$ for any $\eta_0 \in [0, 1]$, differentiating $\eta_0 e^{-z} + (1 - \eta_0)e^z$ with respect to z yields $-\eta_0 e^{-z} + (1 - \eta_0)e^z$; setting this equal to 0 and solving reveals $z^\star(\eta_0) = \frac{1}{2}\ln\left(\frac{\eta_0}{1-\eta_0}\right)$ for $\eta_0 \in (0, 1)$, while $z^\star(0) = -\infty$ and $z^\star(1) = \infty$. Plugging this back in, we have $\ell^\star(\eta_0) = \eta_0 e^{-z^\star(\eta_0)} + (1 - \eta_0)e^{z^\star(\eta_0)} = 2\sqrt{\eta_0(1 - \eta_0)}$. Furthermore, note that if $\eta_0 < 1/2$ and $z > 0$, the derivative $-\eta_0 e^{-z} + (1 - \eta_0)e^z$ of $\eta_0\ell(z) + (1 - \eta_0)\ell(-z)$ is strictly positive; this implies the minimizer of $\eta_0\ell(z) + (1 - \eta_0)\ell(-z)$ over $z \in \bar{\mathbb{R}}$ having $z(2\eta_0 - 1) \leq 0$ is $z = 0$, so that $\ell^\star_-(\eta_0) = \eta_0\ell(0) + (1 - \eta_0)\ell(0) = 1$. Similarly, when $\eta_0 > 1/2$ and $z < 0$, this derivative is strictly negative, so that the minimizer among z with $z(2\eta_0 - 1) \leq 0$ is again $z = 0$, and again $\ell^\star_-(\eta_0) = 1$. For $\eta_0 = 1/2$, we always have $\ell^\star_-(\eta_0) = \ell^\star(\eta_0)$, which in this case equals 1. Thus, $\ell^\star_-$ is constant at 1. In particular, since any $\eta_0 \in [0, 1] \setminus \{1/2\}$ has $2\sqrt{\eta_0(1 - \eta_0)} < 1$, ℓ is classification-calibrated. Furthermore, for any $z \in [0, 1]$, $\tilde{\psi}_\ell(z) = 1 - 2\sqrt{\left(\frac{1+z}{2}\right)\left(\frac{1-z}{2}\right)} = 1 - \sqrt{1 - z^2}$. Differentiating twice with respect to z yields $(1-z^2)^{-1/2} + z^2(1-z^2)^{-3/2}$, which is positive real-valued for $z \in [0,1)$, so that $\tilde{\psi}_\ell$ is convex on $[0,1]$, and therefore $\psi_\ell(z) = \tilde{\psi}_\ell(z) = 1 - \sqrt{1-z^2}$. A Taylor expansion of this function around 0 reveals that it is tightly approximated by $z^2/2$ for small z, and in fact satisfies $\frac{z^2}{2} \leq 1 - \sqrt{1-z^2} \leq \frac{z^2}{2}\left(1+z^2\right) \leq z^2$ for $z \in [0,1]$. Thus, for $\varepsilon \in [0,1]$, $\frac{\varepsilon^{2-\alpha}}{8a} \leq \Psi_\ell(\varepsilon) \leq \frac{\varepsilon^{2-\alpha}}{4a}$, with the lower bound becoming tight for small ε (if $\alpha < 1$).

Example 6.3. $\ell(z) = \max\{1-z, 0\}$, the *hinge loss*.
The hinge loss is another commonly-used loss function, typically associated with margin-based methods such as Support Vector Machines (often accompanied by some type of regularization). Toward calculating $\ell^\star(\eta_0)$, first note that, since any $z > 1$ has $\ell(z) = 0 = \ell(1)$ and $\ell(-z) > 2 = \ell(-1)$, without loss we can always take $z^\star(\eta_0) \in [-1,1]$ (though there sometimes exist other valid choices outside $[-1,1]$ as well). Therefore, $\ell^\star(\eta_0) = \inf_{z\in[-1,1]} \eta_0(1-z) + (1-\eta_0)(1+z) = \inf_{z\in[-1,1]} 1 + (1-2\eta_0)z$; if $\eta_0 < 1/2$, $z \mapsto 1 + (1-2\eta_0)z$ is increasing, so that we can take $z^\star(\eta_0) = -1$, and therefore $\ell^\star(\eta_0) = 1 + (1-2\eta_0)(-1) = 2\eta_0$; if $\eta_0 > 1/2$, $z \mapsto 1 + (1-2\eta_0)z$ is decreasing, so that we can take $z^\star(\eta_0) = 1$, and therefore $\ell^\star(\eta_0) = 1 + (1-2\eta_0) = 2(1-\eta_0)$; otherwise, if $\eta_0 = 1/2$, $z \mapsto 1 + (1-2\eta_0)z$ is constant at 1, so that $\ell^\star(\eta_0) = 1$. In general, we have $\ell^\star(\eta_0) = 2\min\{\eta_0, 1-\eta_0\}$. Next, toward calculating $\ell^\star_-(\eta_0)$, again note that, for the same reasons given above, we may take the minimizer of $\eta_0\ell(z) + (1-\eta_0)\ell(-z)$ over $z \in \bar{\mathbb{R}}$ with $z(2\eta_0 - 1) \leq 0$ to be contained in $[-1,1]$. Furthermore, note that $\eta_0\ell(0) + (1-\eta_0)\ell(0) = 1$, while any $z \in [-1,1]$ with $z(2\eta_0-1) \leq 0$ has $\eta_0\ell(z) + (1-\eta_0)\ell(-z) = 1 + (1-2\eta_0)z \geq 1$. Therefore, every $\eta_0 \in [0,1]$ has $\ell^\star_-(\eta_0) = 1$. In particular, any $\eta_0 \in [0,1] \setminus \{1/2\}$ has $\ell^\star_-(\eta_0) - \ell^\star(\eta_0) = 1 - 2\min\{\eta_0, 1-\eta_0\} > 0$, so that the hinge loss is classification-calibrated. Furthermore, for $z \in [0,1]$, $\tilde{\psi}_\ell(z) = 1 - 2\left(1 - \frac{1+z}{2}\right) = z$; since this is a convex function, we have $\psi_\ell(z) = z$ as well. Thus, for any $\varepsilon \in [0,1]$, $\Psi_\ell(\varepsilon) = \varepsilon/2$.

As for Condition 6.3, for any $x, y \in \bar{\mathcal{Y}}$ with $x \leq y$, $1 \leq x \leq y \Rightarrow |\ell(x) - \ell(y)| = 0$, while $x \leq 1 \leq y \Rightarrow |\ell(x) - \ell(y)| = 1 - x \leq |y - x|$, and $x \leq y \leq 1 \Rightarrow |\ell(x) - \ell(y)| = (1-x) - (1-y) = |y-x|$. In any case, we have $|\ell(x) - \ell(y)| \leq |x - y|$, so that the Lipschitz requirement in Condition 6.3 is satisfied with $L = 1$. However, note that, in light of the above analysis of $z^\star(\eta_0)$ values, for the hinge loss, it is natural to consider sets $\bar{\mathcal{Y}}$ containing $\{-1, 1\}$; but in this case, since $\frac{1}{2}\ell(-1) + \frac{1}{2}\ell(1) - \ell(\frac{1}{2}(-1) + \frac{1}{2}(1)) = 0$, we have $\bar{\delta}_\ell(2) \le 0$; since no values of $C_\ell \in (0, \infty)$ and $r_\ell \in (0, \infty]$ satisfy $C_\ell 2^{r_\ell} \le 0$, we find that ℓ does not satisfy the requirement on the modulus of convexity in Condition 6.3.

As an aside, we note that the fact that $\ell_-^\star(\eta_0) = \ell(0)$ and $\psi_\ell = \tilde{\psi}_\ell$ for all three of these examples is not a coincidence. Indeed, Bartlett, Jordan, and McAuliffe [2006] prove that this will always be the case for any *convex* classification-calibrated loss.

6.4 Passive Learning with a Surrogate Loss

As we did in Chapter 3, we begin by stating the known results for passive learning with a surrogate loss. In this context, the specific passive learning algorithm we will be comparing to is known as *empirical ℓ-risk minimization*, which we denote by ERM_ℓ. Specifically, for any $m \in \mathbb{N}$ and $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^m$, define $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{L}) = \mathrm{argmin}_{f \in \mathcal{F}} \mathrm{R}_\ell(f; \mathcal{L})$. To simplify the discussion, we will suppose the infimum value of $\mathrm{R}_\ell(f; \mathcal{L})$ over $f \in \mathcal{F}$ is actually attained by some $f \in \mathcal{F}$; otherwise, we could let $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{L})$ produce any function $\hat{f}$ with $\mathrm{R}_\ell(\hat{f}; \mathcal{L})$ sufficiently *close* to $\inf_{f \in \mathcal{F}} \mathrm{R}_\ell(f; \mathcal{L})$ (say, with $\mathrm{R}_\ell(\hat{f}; \mathcal{L}) \le \inf_{f \in \mathcal{F}} \mathrm{R}_\ell(f; \mathcal{L}) + 1/m$), without sacrificing any of the results below. As before, we allow ERM_ℓ to break ties arbitrarily in the minimization.

As we did for the analysis of $\mathrm{ERM}(\mathbb{C}, \cdot)$ in Chapter 3 (Lemma 3.1), we begin by stating a basic concentration inequality for excess empirical ℓ-risks. Specifically, the following result is implicit in the work of Giné and Koltchinskii [2006]; an explicit proof of a related result, from the details of which this lemma easily follows, is included in the work of Hanneke and Yang [2012], derived from a wonderfully elegant general result of van der Vaart and Wellner [2011].

Lemma 6.6. There is a constant $c \in (1, \infty)$ such that, for any probability measure P over $\mathcal{X} \times \mathcal{Y}$ satisfying Condition 6.2 with given values b and β, for any set $\mathcal{H}$ of measurable functions $f : \mathcal{X} \to \bar{\mathcal{Y}}$ with $f_{P,\ell}^\star \in \mathcal{H}$

and $\mathrm{vc}(\mathcal{G}_{\mathcal{H}}) \le d_\ell$, any $\gamma \in (0,1)$, and any $m \in \mathbb{N}$, letting

$$U_\ell(m,\gamma) = c\left(\frac{b\left(d_\ell \mathrm{Log}\left(\frac{1}{b}\left(\frac{m}{bd_\ell}\right)^{\frac{\beta}{2-\beta}}\right) + \mathrm{Log}(1/\gamma)\right)}{m}\right)^{\frac{1}{2-\beta}} \wedge \bar{\ell},$$

if $\mathcal{L} = \{(X_1', Y_1'), \ldots, (X_m', Y_m')\} \sim P^m$, then with probability at least $1-\gamma$, $\forall f \in \mathcal{H}$, the following inequalities hold:

$$\mathrm{R}_\ell(f;P) - \mathrm{R}_\ell(f^\star_{P,\ell};P) \le \max\left\{2\left(\mathrm{R}_\ell(f;\mathcal{L}) - \mathrm{R}_\ell(f^\star_{P,\ell};\mathcal{L})\right), U_\ell(m,\gamma)\right\},$$
$$\mathrm{R}_\ell(f;\mathcal{L}) - \inf_{g\in\mathcal{H}} \mathrm{R}_\ell(g;\mathcal{L}) \le \max\left\{2\left(\mathrm{R}_\ell(f;P) - \mathrm{R}_\ell(f^\star_{P,\ell};P)\right), U_\ell(m,\gamma)\right\}.$$

The constant c in Lemma 6.6 may depend on ℓ via $\bar{\ell}$, which (as mentioned above) we are treating as a numerical constant here to simplify the discussion; the interested reader is referred to the work of Hanneke and Yang [2012] for an explicit description of this dependence. For completeness, also define $U_\ell(0,\gamma) = \bar{\ell}$. The factor of $\mathrm{Log}\left(\frac{1}{b}\left(\frac{m}{bd_\ell}\right)^{\frac{\beta}{2-\beta}}\right)$ in the definition of $U_\ell(m,\gamma)$ can be refined based on a generalization of the disagreement coefficient, so that $U_\ell(m,\gamma)$ has a form similar to that of $U(m,\gamma)$; for simplicity, we do not present the details of this refinement here, again referring the interested reader to the work of Hanneke and Yang [2012] for this.

As before, it is fairly straightforward to apply Lemma 6.6 to arrive at a bound on the label complexity achieved by $\mathrm{ERM}_\ell(\mathcal{F},\cdot)$. Specifically, we may simply solve the bound in Lemma 6.6 (applied to $\mathcal{H} = \mathcal{F}$ and $P = \mathcal{P}_{XY}$) for a value of m sufficiently large to guarantee $U_\ell(m,\delta) \le \Psi_\ell(\varepsilon)$, so that (with probability at least $1-\delta$), the empirical ℓ-risk minimizer $\hat{f}$ has $\mathrm{R}_\ell(\hat{f}) - \mathrm{R}_\ell(f^\star_\ell) \le \Psi_\ell(\varepsilon)$; Lemma 6.5 then implies $\mathrm{er}(\hat{f}) - \mathrm{er}(f^\star) \le \varepsilon$. Specifically, the following implication, analogous to (3.1), is helpful in deriving such a result; it follows by simple algebra. For some constant $c' \in [1,\infty)$ (depending only on $\bar{\ell}$), for any $m \in \mathbb{N}$, $\varepsilon > 0$, and $\gamma \in (0,1)$,

$$m \ge c' b\varepsilon^{\beta-2}\left(d_\ell \mathrm{Log}\left(1/(b\varepsilon^\beta)\right) + \mathrm{Log}\left(1/\gamma\right)\right) \implies U_\ell(m,\gamma) \le \varepsilon. \quad (6.5)$$

With this in hand, we have the following theorem; again, this result is implicit in the work of Giné and Koltchinskii [2006], and there is an explicit proof in the work of Hanneke and Yang [2012] (with slightly refined logarithmic factors compared to the result presented here).

Theorem 6.7. If ℓ is classification-calibrated, then the passive learning algorithm $\mathrm{ERM}_\ell(\mathcal{F}, \cdot)$ achieves a label complexity Λ such that, for any distribution $\mathcal{P}_{XY}$ satisfying Condition 6.2 with given values b and β, satisfying Condition 2.3 with given values a and α, and having $f_\ell^\star \in \mathcal{F}$, $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} \left(d_\ell \mathrm{Log}\left(1/(b\Psi_\ell(\varepsilon)^\beta)\right) + \mathrm{Log}(1/\delta)\right).$$

6.5 Active Learning with a Surrogate Loss

Inspired by results such as Theorem 6.7 for passive learning, it is interesting to consider using surrogate losses in the design of active learning algorithms as well. The aim is to design computationally efficient methods, still capable of achieving the same types of improvements (in this case, compared to Theorem 6.7) that we found for methods that directly optimize the 0-1 loss (namely, RobustCAL). Toward this end, this section discusses a modification of RobustCAL to replace optimizations involving empirical error rates with optimizations involving empirical ℓ-risks, for an arbitrary classification-calibrated loss ℓ. We indeed find that such a method achieves label complexity improvements over Theorem 6.7, of a similar type to those proven previously for RobustCAL relative to Theorem 3.4.

6.5.1 Modifying RobustCAL to Use a Surrogate Loss

Lemma 6.6 offers a direction for working with surrogate losses in active learning. It suggests the possibility of replacing the empirical error rates in RobustCAL with empirical ℓ-risks, while replacing $U(m,\gamma)$ with $U_\ell(m,\gamma)$ to compensate. This is indeed the strategy explored by Hanneke and Yang [2012], and which we will now discuss in detail. The key observation that enables us to make this substitution is that, since we are only interested in using ℓ as a *surrogate loss*, and beyond this have no genuine interest in optimizing $\mathrm{R}_\ell(\cdot)$, once we have identified $\mathrm{sign}(f_\ell^\star(x))$ for points x in a given region $S \subset \mathcal{X}$, we need not concern ourselves with optimizing the ℓ-risk any further on the set S, so that we can focus our efforts on the remaining region $\mathcal{X} \setminus S$. Thus, even though we are using the empirical ℓ-risk in the algorithm, we do not require any guarantee on the ℓ-risk of the function produced at the conclusion of the algorithm: only a guarantee on its *error rate*. With this observation in mind, the above-described algorithm (with a few technical modifications) is stated formally as follows (where, as it was in RobustCAL, $\delta_m = \delta/(\log_2(2m))^2$ for $\delta \in (0,1)$ and $m \in \mathbb{N}$).

Algorithm: $\mathbf{RobustCAL}_\delta^\ell(n)$
0. $m \leftarrow 1$, $Q \leftarrow \{\}$, $V \leftarrow \mathcal{F}$, $t \leftarrow 0$
1. While $t < n$ and $m < 2^n$
2. $\quad m \leftarrow m + 1$
3. $\quad$ If $X_m \in \mathrm{DIS}(V)$
4. $\quad\quad$ Request the label Y_m; let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$, $t \leftarrow t + 1$
5. $\quad$ If $\log_2(m) \in \mathbb{N}$
6. $\quad\quad V \leftarrow \left\{ f \in V : \left(\mathrm{R}_\ell(f;Q) - \inf\limits_{g \in V} \mathrm{R}_\ell(g;Q)\right)|Q| \leq U_\ell(\frac{m}{2}, \delta_m)\frac{m}{2} \right\}$
$\quad\quad Q \leftarrow \{\}$
7. Return $\hat{h} = \mathrm{sign}(\hat{f})$ for any $\hat{f} \in V$

As was the case for CAL and the original RobustCAL, there are (at least) two ways to describe the algorithm. The above is a particularly convenient description of $\mathrm{RobustCAL}_\delta^\ell$ for the purpose of theoretical analysis, since it explicitly represents the set V of classifiers still under consideration at any given time. This simplifies the notation in the label complexity analysis. However, in practice one would typically only maintain this set *implicitly* as a set of constraints, so that running the algorithm essentially involves solving a sequence of straightforward constraint satisfaction and constrained optimization problems. Specifically, the following algorithm behaves identically to that stated above, without explicitly representing the set V.

Algorithm: $\mathbf{RobustCAL}_{\delta}^{\ell}(n)$
0. $m \leftarrow 1$, $i \leftarrow 1$, $Q_i \leftarrow \{\}$
1. While $\sum_{j=1}^{i} |Q_j| < n$ and $m < 2^n$
2. $\quad m \leftarrow m+1$
3. $\quad$ If $\forall y \in \mathcal{Y}$, $\exists f \in \mathcal{F}$ s.t. $yf(X_m) \geq 0$ (with $yf(X_m) > 0$ if $y = -1$) and $\forall j < i, (\mathrm{R}_{\ell}(f; Q_j) - \mathrm{R}_j^*)|Q_j| \leq U_{\ell}(2^{j-1}, \delta_{(2^j)})2^{j-1}$
4. $\quad\quad$ Request the label Y_m; let $Q_i \leftarrow Q_i \cup \{(X_m, Y_m)\}$
5. $\quad$ If $\log_2(m) \in \mathbb{N}$
6. $\quad\quad \mathrm{R}_i^* \leftarrow \inf \Big\{ \mathrm{R}_{\ell}(f; Q_i) : f \in \mathcal{F}$ and $\forall j < i, (\mathrm{R}_{\ell}(f; Q_j) - \mathrm{R}_j^*)|Q_j| \leq U_{\ell}(2^{j-1}, \delta_{(2^j)})2^{j-1} \Big\}$
$\quad\quad i \leftarrow i+1$
$\quad\quad Q_i \leftarrow \{\}$
7. Return $\hat{h} = \mathrm{sign}(\hat{f})$ for any $\hat{f} \in \mathcal{F}$ s.t. $\forall j < i, (\mathrm{R}_{\ell}(\hat{f}; Q_j) - \mathrm{R}_j^*)|Q_j| \leq U_{\ell}(2^j, \delta_{(2^j)})2^j$

In the special case of $\mathcal{F} = \mathbb{C}$ and $\ell = \ell_{01}$, the 0-1 loss itself, the above method is quite similar to the original RobustCAL algorithm stated in Chapter 5. The main difference is that we are resetting the set Q in Step 6, which is done for the technical reason that we want to apply Lemma 6.6 with a set $\mathcal{H}$ that depends on V, and therefore depends on $\mathcal{Z}_{m/2}$; thus, for the lemma to be applicable (without modification), we need to be applying it only with the most-recent $m/2$ data points, so that V and $\mathcal{L}$ are independent. The other subtle difference is that the quantity U_ℓ is slightly looser than U (though see the discussion above about this), and furthermore is based on the values b and β from Condition 6.2, which in the special case of the 0-1 loss, corresponds to the stronger condition (2.1) (or (2.2)), rather than the weaker Condition 2.3 that defines the values a and α in the quantity U used in the original RobustCAL algorithm.

Replacing $U_\ell(m, \gamma)$ with a data-dependent estimator: As was the case for the original RobustCAL, it is possible to replace the $U_\ell(m, \gamma)$ quantities with appropriate data-dependent values, which have no direct dependence on $\mathcal{P}_{XY}$ (other than the data), and which are upper-bounded by $U_\ell(m, \gamma)$ with high probability [Hanneke and Yang, 2012];

the results below remain valid, even with this replacement. To focus on the details most essential to the label complexity analysis, and avoid complicating the analysis with the details of relating these data-dependent quantities to $U_\ell(m,\gamma)$, we do not discuss this substitution here; the interested reader is referred to the work of Hanneke and Yang [2012] for those details.

6.5.2 General Label Complexity Analysis

The proof of Theorem 5.4 requires surprisingly few modifications to obtain a corresponding general result for RobustCAL$_\delta^\ell$, holding for any classification-calibrated ℓ. The idea is to apply Lemma 6.6 to the set $\mathcal{H}$ of functions $f = g\mathbb{1}_{\mathrm{DIS}(V)} + f_\ell^\star \mathbb{1}_{\mathcal{X}\setminus\mathrm{DIS}(V)}$ for $g \in V$. If $f_\ell^\star \in V$, then $f_\ell^\star \in \mathcal{H}$ as well. Thus, for any m with $\log_2(m) \in \mathbb{N}$, letting

$$\mathcal{L}_m = \{(X_i, Y_i)\}_{i=\frac{m}{2}+1}^{m},$$

Lemma 6.6 implies that, with probability $1-\delta_m$, $\mathrm{R}_\ell(f_\ell^\star;\mathcal{L}_m) - \inf_{g\in\mathcal{H}} \mathrm{R}_\ell(g;\mathcal{L}_m) \le U_\ell(m/2,\delta_m)$ and $\forall f \in \mathcal{H}$, $\mathrm{R}_\ell(f) - \mathrm{R}_\ell(f_\ell^\star) \le \max\{2(\mathrm{R}_\ell(f;\mathcal{L}_m) - \mathrm{R}_\ell(f_\ell^\star;\mathcal{L}_m)), U_\ell(m/2,\delta_m)\}$. Since the set Q upon reaching Step 6 is precisely the subsequence of points (X_t,Y_t) in $\mathcal{L}_m$ with $X_t \in \mathrm{DIS}(V)$, and any (X_t,Y_t) in $\mathcal{L}_m$ with $X_t \notin \mathrm{DIS}(V)$ has $f(X_t) = f_\ell^\star(X_t)$ for every $f \in \mathcal{H}$, we have $\forall f \in \mathcal{H}$, $(\mathrm{R}_\ell(f;\mathcal{L}_m) - \mathrm{R}_\ell(f_\ell^\star;\mathcal{L}_m))m/2 = (\mathrm{R}_\ell(f;Q) - \mathrm{R}_\ell(f_\ell^\star;Q))|Q|$. Combined with the above observations, we have $(\mathrm{R}_\ell(f_\ell^\star;Q) - \inf_{g\in V}\mathrm{R}_\ell(g;Q))\,|Q| = (\mathrm{R}_\ell(f_\ell^\star;\mathcal{L}_m) - \inf_{g\in\mathcal{H}}\mathrm{R}_\ell(g;\mathcal{L}_m))\,m/2 \le U_\ell(m/2,\delta_m)m/2$, so that $f_\ell^\star$ will remain in V after the update in Step 6. Furthermore, for every $g \in V$ that survives the update in Step 6, letting f be the corresponding function in $\mathcal{H}$, we have $\mathrm{R}_\ell(f) - \mathrm{R}_\ell(f_\ell^\star) \le \max\{2(\mathrm{R}_\ell(f;\mathcal{L}_m) - \mathrm{R}_\ell(f_\ell^\star;\mathcal{L}_m)), U_\ell(m/2,\delta_m)\} = \max\left\{2(\mathrm{R}_\ell(g;Q) - \mathrm{R}_\ell(f_\ell^\star;Q))\frac{2|Q|}{m}, U_\ell(m/2,\delta_m)\right\} \le 2U_\ell(m/2,\delta_m)$. Applying this argument inductively, each time we update the set V, we maintain the invariants that $f_\ell^\star \in V$ and $\forall g \in V$, the corresponding function $f \in \mathcal{H}$ has $\mathrm{R}_\ell(f) - \mathrm{R}_\ell(f_\ell^\star) \le 2U_\ell(m/2,\delta_m)$. We can then use this guarantee to arrive at a label complexity bound by simply taking m sufficiently large to guarantee $2U_\ell(m/2,\delta_m) \le \Psi_\ell(\varepsilon)$, in which case Lemma 6.5 implies every $f \in \mathcal{H}$ has $\mathrm{er}(f) - \mathrm{er}(f^\star) \le \varepsilon$; since the function $g \in V$ corresponding to f has $\mathrm{sign}(g) = \mathrm{sign}(f)$, this also means $\mathrm{er}(g) - \mathrm{er}(f^\star) \le \varepsilon$. The details of this argument are formalized in the proof of Theorem 6.8 below, due to Hanneke and Yang [2012] (though the original result has slightly refined logarithmic factors compared to those in the theorem presented here).

Theorem 6.8. Suppose ℓ is classification-calibrated. For any $\delta \in (0,1)$, $\mathrm{RobustCAL}_\delta^\ell$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ satisfying Condition 2.3 with given values a and α, satisfying Condition 6.2 with given values b and β, and with $f_\ell^\star \in \mathcal{F}$, $\forall \varepsilon \in (0,1)$,

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim$$
$$\frac{\theta(a\varepsilon^\alpha)\, a\varepsilon^\alpha b}{\Psi_\ell(\varepsilon)^{2-\beta}}\left(d_\ell \mathrm{Log}\!\left(\frac{1}{b\Psi_\ell(\varepsilon)^\beta}\right) + \mathrm{Log}\!\left(\frac{\mathrm{Log}\!\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right)\mathrm{Log}\!\left(\frac{1}{\Psi_\ell(\varepsilon)}\right).$$

Proof. The proof is essentially similar to the proof of Theorem 5.4, except slightly more involved to handle the use of ℓ in place of ℓ_{01}.

Fix $\varepsilon, \delta \in (0,1)$, and consider running $\mathrm{RobustCAL}_\delta^\ell$ with budget argument $n \in \mathbb{N}$ greater than

$$c''\frac{\theta(a\varepsilon^\alpha)\, a\varepsilon^\alpha b}{\Psi_\ell(\varepsilon)^{2-\beta}}\left(d_\ell \mathrm{Log}\!\left(\frac{1}{b\Psi_\ell(\varepsilon)^\beta}\right) + \mathrm{Log}\!\left(\frac{\mathrm{Log}\!\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right)\mathrm{Log}\!\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)$$

for an appropriate constant c'' (indicated by the analysis below). Let $M \subseteq \{0,\ldots,2^n\}$ denote the set of values of m obtained during the execution. Let $V_1 = \mathcal{F}$ and $Q_1 = \emptyset$, and for each $m \in M$ with $\log_2(m) \in \mathbb{N}$, let Q_m denote the value of Q upon reaching Step 5 with that value of m, and let V_m denote the value of V upon completion of Step 6 (i.e., after the update). Also, for each $f \in \mathcal{F}$ and $m \in M$ with $\log_2(m) \in \mathbb{N} \cup \{0\}$, define $f_m = f\mathbb{1}_{\mathrm{DIS}(V_m)} + f_\ell^\star \mathbb{1}_{\mathcal{X}\setminus\mathrm{DIS}(V_m)}$, and then define the set $\mathcal{H}_m = \{f_m : f \in \mathcal{F}\}$. Also let $\mathcal{L}_m = \{(X_i,Y_i)\}_{i=\frac{m}{2}+1}^m$ for m with $\log_2(m) \in \mathbb{N}$, as defined above.

First note that, for any $m \in M$ with $\log_2(m) \in \mathbb{N}\cup\{0\}$, $t \in \mathbb{N}$, and any sequences of points $\{(x_i,y_i,z_i)\}_{i=1}^t \in (\mathcal{X}\times\mathcal{Y}\times\mathbb{R})^t$ shattered by $\mathcal{G}_{\mathcal{H}_m}$, it must be that none of the x_i are in $\mathcal{X}\setminus\mathrm{DIS}(V_m)$ (since the functions $f_m \in \mathcal{H}_m$ all agree on values in $\mathcal{X}\setminus\mathrm{DIS}(V_m)$); therefore $\{(f_m(x_1),\ldots,f_m(x_t)) : f_m \in \mathcal{H}_m\} = \{(f(x_1),\ldots,f(x_t)) : f \in \mathcal{F}\}$, so that $\mathcal{G}_{\mathcal{F}}$ also shatters these t points; this implies $\mathrm{vc}(\mathcal{G}_{\mathcal{H}_m}) \leq d_\ell$. Furthermore, note that $f_\ell^\star \in \mathcal{H}_m$. Therefore, applying Lemma 6.6 conditioned on $V_{m/2}$ (which is independent of $\mathcal{L}_m$), together with the law of total probability (to integrate out the $V_{m/2}$ variable) and a union bound (over values of m with $\log_2(m) \in \mathbb{N}$), we have that, on an event E_δ of probability at least $1 - \sum_{i=1}^{\infty} \frac{\delta}{(1+i)^2} > 1 - 2\delta/3$, every $m \in M$ with $\log_2(m) \in \mathbb{N}$ has

$$\mathrm{R}_\ell(f_\ell^\star; \mathcal{L}_m) - \inf_{g \in \mathcal{H}_{m/2}} \mathrm{R}_\ell(g; \mathcal{L}_m) \leq U_\ell(m/2, \delta_m), \tag{6.6}$$

and $\forall f_{m/2} \in \mathcal{H}_{m/2}$,

$$\mathrm{R}_\ell(f_{m/2}) - \mathrm{R}_\ell(f_\ell^\star) \leq \max\Big\{2(\mathrm{R}_\ell(f_{m/2}; \mathcal{L}_m) - \mathrm{R}_\ell(f_\ell^\star; \mathcal{L}_m)), U_\ell(m/2, \delta_m)\Big\}. \tag{6.7}$$

In particular, since (as noted above) every $m \in M$ with $\log_2(m) \in \mathbb{N}$, and every $f, g \in V_{m/2}$, have

$$(\mathrm{R}_\ell(f; Q_m) - \mathrm{R}_\ell(g; Q_m))\,|Q_m| = \Big(\mathrm{R}_\ell(f_{m/2}; \mathcal{L}_m) - \mathrm{R}_\ell(g_{m/2}; \mathcal{L}_m)\Big)\, m/2, \tag{6.8}$$

combined with (6.6) this implies that if $f_\ell^\star \in V_{m/2}$, then

$$\left(\mathrm{R}_\ell(f_\ell^\star; Q_m) - \inf_{g \in V_{m/2}} \mathrm{R}_\ell(g; Q_m)\right) |Q_m| \leq U_\ell(m/2, \delta_m) m/2,$$

so that $f_\ell^\star \in V_m$ as well. Furthermore, this fact, combined with (6.7), also implies that if $f_\ell^\star \in V_{m/2}$ for some $m \in M$ with $\log_2(m) \in \mathbb{N}$, then (by definition of V_m in Step 6) on the event E_δ, every $f \in V_m$ has

$$\begin{aligned}
\mathrm{R}_\ell(f_{m/2}) &- \mathrm{R}_\ell(f_\ell^\star) \\
&\leq \max\left\{2(\mathrm{R}_\ell(f_{m/2}; \mathcal{L}_m) - \mathrm{R}_\ell(f_\ell^\star; \mathcal{L}_m)), U_\ell(m/2, \delta_m)\right\} \\
&= \max\left\{2(\mathrm{R}_\ell(f; Q_m) - \mathrm{R}_\ell(f_\ell^\star; Q_m))\frac{2|Q_m|}{m}, U_\ell(m/2, \delta_m)\right\} \\
&\leq 2U_\ell(m/2, \delta_m).
\end{aligned}$$

Thus, since $f_\ell^\star \in V_1 = \mathcal{F}$, by induction we have that, on the event E_δ, every $m \in M$ with $\log_2(m) \in \mathbb{N}$ has $f_\ell^\star \in V_m$, and every $f \in V_m$ has $\mathrm{R}_\ell(f_{m/2}) - \mathrm{R}_\ell(f_\ell^\star) \leq 2U_\ell(m/2, \delta_m)$.

Now let $i_\varepsilon = \lceil \log_2(2/\Psi_\ell(\varepsilon)) \rceil$, define $I = \{0, \ldots, i_\varepsilon\}$, and for $i \in I \setminus \{0\}$ let $\varepsilon_i = 2^{-i}$; also let $\varepsilon_0 = \bar{\ell}/2 = \left(1 \vee \sup_{z \in \bar{\mathcal{Y}}} \ell(z)\right)/2$. Let $m_0 = 1$, and for c' as in (6.5), for each $i \in I \setminus \{0\}$, define

$$m_i' = 4c'b\varepsilon_i^{\beta-2}\left(d_\ell \mathrm{Log}\left(\frac{1}{b\varepsilon_i^\beta}\right) + \mathrm{Log}\left(\frac{4\log_2(4c'b/\varepsilon_i)}{\delta}\right)\right)$$

and $m_i = 2^{1+\lceil \log_2(m_i') \rceil}$. One can easily check that, for $i \in I \setminus \{0\}$, $m_i/2$ satisfies the condition of (6.5) with $\gamma = \delta_{m_i}$, so that $U_\ell(m_i/2, \delta_{m_i}) \le \varepsilon_i$ for all $i \in I \setminus \{0\}$. In particular, for every $i \in I \setminus \{0\}$ with $m_i \in M$, the conclusions of the previous paragraph imply that, on the event E_δ,

$$\forall f \in V_{m_i}, \mathrm{R}_\ell(f_{m_i/2}) - \mathrm{R}_\ell(f_\ell^\star) \le 2U_\ell(m_i/2, \delta_{m_i}) \le 2\varepsilon_i. \tag{6.9}$$

Furthermore, since $V_{m_i} \subseteq V_{m_i/2}$, $\mathrm{sign}(f_{m_i/2}) = \mathrm{sign}(f)$ for every $f \in V_{m_i}$, so that $\mathrm{sign}(f_{m_i/2}) \in \mathbb{C}$; thus, (6.9) and Lemma 6.5 imply that, on E_δ, every $f \in V_{m_i}$ has $\mathrm{er}(f_{m_i/2}) - \mathrm{er}(f^\star) \le \Psi_\ell^{-1}(2\varepsilon_i)$, where Ψ_ℓ^{-1} is the inverse of Ψ_ℓ, which is well-defined in this context since Ψ_ℓ is continuous and strictly increasing for classification-calibrated ℓ. Since $\mathrm{sign}(f) = \mathrm{sign}(f_{m_i/2})$, we have $\mathrm{er}(f) = \mathrm{er}(f_{m_i/2})$, so that $er(f) - \mathrm{er}(f^\star) \le \Psi_\ell^{-1}(2\varepsilon_i)$ as well. Furthermore, since $\Psi_\ell(1) = a\psi_\ell(1/(2a)) \le \psi_\ell(1/2) \le \tilde{\psi}_\ell(1/2) \le \bar{\ell} = 2\varepsilon_0$, we have $\Psi_\ell^{-1}(2\varepsilon_0) \ge 1$; thus, since every $f \in \mathcal{F}$ trivially has $\mathrm{er}(f) - \mathrm{er}(f^\star) \le 1$, we have established that, on the event E_δ, every $i \in I$ with $m_i \in M$ satisfies

$$\forall f \in V_{m_i}, \mathrm{er}(f) - \mathrm{er}(f^\star) \le \Psi_\ell^{-1}(2\varepsilon_i). \tag{6.10}$$

The remainder of the proof focuses on analyzing the number of labels requested between updates of V, with the intention of showing that $m_{i_\varepsilon} \in M$ with high probability; as we will see below, (6.10) implies that this will be sufficient to complete the proof. We can express the number of labels requested while $m \le m_{i_\varepsilon}$ as

$$\sum_{t=0}^{\log_2(m_{i_\varepsilon}/2)} \sum_{m=2^t+1}^{\min\{2^{t+1}, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{2^t})}(X_m) \le \sum_{i=1}^{i_\varepsilon} \sum_{m=m_{i-1}+1}^{\min\{m_i, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{m_{i-1}})}(X_m).$$

Now note that, by (6.10), on the event E_δ, for $i \in I \setminus \{0\}$, $\mathrm{DIS}(V_{m_{i-1}}) \subseteq \mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)\right)\right)$, so that the above summation is at most

$$\sum_{i=1}^{i_\varepsilon} \sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}(2\varepsilon_{i-1})\right)\right)}(X_m).$$

This is a sum of independent Bernoulli random variables, so that a Chernoff bound implies that, on an event E'_δ of probability at least $1-\delta/3$, the value of the sum is at most

$$\log_2(3/\delta) + 2e \sum_{i=1}^{i_\varepsilon} (m_i - m_{i-1})\, \mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)\right)\right)\right). \tag{6.11}$$

Condition 2.3 implies that for $i \in I \setminus \{0\}$, $\mathbb{C}\left(\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)\right) \subseteq \mathrm{B}\left(f^\star, a\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)^\alpha\right)$, so that $\mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)\right)\right)\right) \leq \theta\left(a\Psi_\ell^{-1}(2\varepsilon_{i-1})^\alpha\right) a\Psi_\ell^{-1}(2\varepsilon_{i-1})^\alpha$. Combined with the definition of m_i, we have that for every $i \in I \setminus \{0\}$,

$$m_i \mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}\left(2\varepsilon_{i-1}\right)\right)\right)\right) \lesssim \frac{\theta\left(a\Psi_\ell^{-1}(2\varepsilon_{i-1})^\alpha\right) a\Psi_\ell^{-1}(2\varepsilon_{i-1})^\alpha b}{\varepsilon_i^{2-\beta}} \left(d_\ell \mathrm{Log}\left(\frac{1}{b\varepsilon_i^\beta}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\varepsilon_i}\right)}{\delta}\right)\right). \tag{6.12}$$

Plugging this into (6.11), and using basic properties of $\theta(\cdot)$ (namely, Theorem 7.1 and Corollary 7.2 from Chapter 7 below), combined with the fact that every $i \in I$ has $\varepsilon_i > \Psi_\ell(\varepsilon)/4$, we find that (6.11) is

$$\lesssim \frac{\theta(a\varepsilon^\alpha)\, a\varepsilon^\alpha b}{\Psi_\ell(\varepsilon)^{2-\alpha}} \left(d_\ell \mathrm{Log}\left(\frac{1}{b\Psi_\ell(\varepsilon)^\beta}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right) \mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right).$$

For an appropriate choice of the constant c'', the budget n is larger than this, and furthermore $m_{i_\varepsilon} < 2^n$. Thus, we have proven that, on $E_\delta \cap E'_\delta$, we have $m_{i_\varepsilon} \in M$, so that $\hat{f} \in V_{m_{i_\varepsilon}}$. In particular, this means that (6.10) implies $\mathrm{er}(\hat{f}) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}(2\varepsilon_{i_\varepsilon}) \leq \Psi_\ell^{-1}(\Psi_\ell(\varepsilon)) = \varepsilon$. Noting that $\hat{h} = \mathrm{sign}(\hat{f})$ implies $\mathrm{er}(\hat{h}) = \mathrm{er}(\hat{f})$, we have $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. Furthermore, by a union bound, the event $E_\delta \cap E'_\delta$ has probability at least $1-\delta$. Noting that the above sufficient condition on n matches the form of the bound from the theorem statement completes the proof. $\square$

Aside from logarithmic factors, the label complexity bound in Theorem 6.8 is essentially a factor of $\theta(a\varepsilon^{\alpha})a\varepsilon^{\alpha}$ smaller than the bound in Theorem 6.7 for $\mathrm{ERM}_{\ell}(\mathcal{F},\cdot)$. Note that this is the same *type* of improvement we found in Theorem 5.4 for RobustCAL. We can obtain more-concrete results by plugging in the various quantities for the surrogate losses given above in Section 6.3. For instance, when ℓ is the quadratic loss, we find an asymptotic dependence on ε of $O\left(\theta(\varepsilon^{\alpha})\varepsilon^{2(\alpha-1)}(\mathrm{Log}(1/\varepsilon))^2\right)$ (using the value of β indicated by Lemma 6.4). This is roughly similar in form to the results obtained previously for $\mathrm{RobustCAL}_{\delta}$ (i.e., without using a surrogate loss); however, the conditions leading to validity of this result for $\mathrm{RobustCAL}_{\delta}^{\ell}$ are significantly stronger, since we require $f_{\ell}^{\star} \in \mathcal{F}$; recall that, in the case of the quadratic loss, this is equivalent to the assumption that $(2\eta(\cdot)-1) \in \mathcal{F}$.

The logarithmic factors in Theorem 6.8 can be improved in some cases [Hanneke and Yang, 2012]. As was the case for Theorem 6.7, the $d_{\ell}\mathrm{Log}(1/(b\Psi_{\ell}(\varepsilon))^{\beta})$ term can be refined based on a generalization of the disagreement coefficient (though a slightly different generalization than the one alluded to above for Theorem 6.7); the interested reader is referred to the work of Hanneke and Yang [2012] for the definition of this generalization, and its use in refining this logarithmic factor. Additionally, as was true of the bound in Theorem 5.4, by a careful treatment of the summations in the proof, the final $\mathrm{Log}(1/\Psi_{\ell}(\varepsilon))$ factor can be eliminated (replaced by a constant factor) when either $\alpha+\beta<2$ or $\theta(\varepsilon)=\Theta(\varepsilon^{-\kappa})$ for some $\kappa \in (0,1]$; Furthermore, under these same conditions, it is also possible to eliminate the $\mathrm{LogLog}(b/\Psi_{\ell}(\varepsilon))$ term (replacing it with a constant) by using a modified ($\mathcal{P}_{XY}$-dependent) definition of the δ_m values [see Hanneke and Yang, 2012].

6.5.3 Label Complexity Analysis for Smooth Convex Losses

In the special case of losses ℓ satisfying Condition 6.3, we can get a sometimes-tighter result by using a threshold in Step 6 based on the *conditional* distribution of (X,Y) given that $X \in \mathrm{DIS}(V)$. Specifically, let b and β be the values indicated by Lemma 6.4, and then let $U_{\ell}(m,\gamma)$ be defined as in Lemma 6.6 with these values of b and β. Then consider replacing Step 6 in $\mathrm{RobustCAL}_{\delta}^{\ell}$ with the following alternative.

$$6'.\quad V \leftarrow \left\{ f \in V : \mathrm{R}_{\ell}(f;Q) - \inf_{g \in V} \mathrm{R}_{\ell}(g;Q) \leq U_{\ell}(|Q|, \delta_m) \right\}; Q \leftarrow \{\}$$

The idea is that, since Lemma 6.4 indicates Condition 6.2 will also be satisfied under the *conditional* distribution given $X \in \mathrm{DIS}(V)$, which is precisely the distribution governing the samples in Q, we can directly apply Lemma 6.6 to the set Q and this conditional distribution; by the same reasoning as above, this will (with high probability) never remove $f_{\ell}^{\star}$ from V, but will more aggressively prune down the set V compared to the original update from Step 6 (assuming the same values of b and β are used in both cases). Formally, this modification leads to the following result, originally due to Hanneke and Yang [2012].

Theorem 6.9. Suppose ℓ is classification-calibrated, and satisfies Condition 6.3, and let b and β be the values indicated by Lemma 6.4. For any $\delta \in (0,1)$, if we replaced Step 6 in $\mathrm{RobustCAL}_{\delta}^{\ell}$ with Step $6'$ described above, the resulting algorithm achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ satisfying Condition 2.3 with given values a and α, and with $f_{\ell}^{\star} \in \mathcal{F}$, $\forall \varepsilon \in (0,1)$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim$$
$$b \left(\frac{\theta(a\varepsilon^{\alpha}) a\varepsilon^{\alpha}}{\Psi_{\ell}(\varepsilon)} \right)^{2-\beta} \left(d_{\ell} \mathrm{Log}\left(\frac{1}{b} \left(\frac{\theta(a\varepsilon^{\alpha}) a\varepsilon^{\alpha}}{\Psi_{\ell}(\varepsilon)} \right)^{\beta} \right) \right.$$
$$\left. + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\Psi_{\ell}(\varepsilon)} \right)}{\delta} \right) \right) \mathrm{Log}\left(\frac{1}{\Psi_{\ell}(\varepsilon)} \right).$$

Proof. Fix any $\varepsilon, \delta \in (0,1)$. For each $i \in \mathbb{N} \cup \{0\}$, let $\varepsilon_i = \bar{\ell} 2^{-i}$, and $p_i = \mathcal{P}\left(\mathrm{DIS}\left(\mathbb{C}\left(\Psi_{\ell}^{-1}(\varepsilon_i) \right) \right) \right)$. Let $m_0 = 1$, and for each $i \in \mathbb{N}$, define

$$m_i' = c'' \frac{b\left((p_{i-1} \vee \varepsilon_i)^{1-\beta} \right)}{\varepsilon_i^{2-\beta}} \left(d_{\ell} \mathrm{Log}\left(\frac{1}{b} \left(\frac{\bar{\ell} p_{i-1}}{\varepsilon_i} \right)^{\beta} \right) + \mathrm{Log}\left(\frac{\log_2\left(\frac{b\bar{\ell}}{\varepsilon_i} \right)}{\delta} \right) \right),$$

for a constant $c'' \in (1, \infty)$ indicated by the analysis below, and let $m_i = \max\left\{ 2m_{i-1}, 2^{1+\lceil \log_2(m_i') \rceil} \right\}$. Let $i_{\varepsilon} = \lceil \log_2(\bar{\ell}/\Psi_{\ell}(\varepsilon)) \rceil$, and let $I =$

$\{1, \ldots, i_\varepsilon\}$. Now consider running $\mathrm{RobustCAL}_\delta^\ell$ with budget argument $n \in \mathbb{N}$ satisfying

$$n \geq c_0 b \left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} \left(d_\ell \mathrm{Log}\left(\frac{1}{b}\left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^\beta\right) + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right),$$

for an appropriate numerical constant $c_0 > 8c''$ (indicated by the analysis below). Note that for each $i \in I$, Condition 2.3 and the definition of the disagreement coefficient imply $p_{i-1} \leq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f^\star, a\Psi_\ell^{-1}(\varepsilon_{i-1})^\alpha\right)\right)\right) \leq \theta\left(a\Psi_\ell^{-1}(\varepsilon_{i-1})^\alpha\right)a\Psi_\ell^{-1}(\varepsilon_{i-1})^\alpha$. Furthermore, note that $\varepsilon_{i-1} \geq \varepsilon_{i_\varepsilon - 1} > \Psi_\ell(\varepsilon)$, so that monotonicity of Ψ_ℓ^{-1} and θ imply $\theta\left(a\Psi_\ell^{-1}(\varepsilon_{i-1})^\alpha\right) \leq \theta\left(a\Psi_\ell^{-1}\left(\varepsilon_{i_\varepsilon-1}\right)^\alpha\right) \leq \theta\left(a\Psi_\ell^{-1}\left(\Psi_\ell(\varepsilon)\right)^\alpha\right) = \theta\left(a\varepsilon^\alpha\right)$. For any $x, t \in (0, \infty)$ with $t \geq 1$, we have $\Psi_\ell(xt) = a(xt)^\alpha\psi_\ell\left(\frac{(xt)^{1-\alpha}}{2a}\right) \geq a(xt)^\alpha\psi_\ell\left(\frac{(xt)^{1-\alpha}}{2a}\frac{1}{t^{1-\alpha}}\right)t^{1-\alpha} = \Psi_\ell(x)t$. Thus, for $z = \Psi_\ell^{-1}(\varepsilon_{i-1})/\varepsilon_i$, we have $\varepsilon_{i-1} = \Psi_\ell(\varepsilon_i z) = \Psi_\ell(\varepsilon_{i_\varepsilon+1}z(\varepsilon_i/\varepsilon_{i_\varepsilon+1})) \geq \Psi_\ell(\varepsilon_{i_\varepsilon+1}z)(\varepsilon_i/\varepsilon_{i_\varepsilon+1}) = \Psi_\ell(\varepsilon_{i_\varepsilon+1}z)(\varepsilon_{i-1}/\varepsilon_{i_\varepsilon})$, so that $\Psi_\ell(\varepsilon_{i_\varepsilon+1}z) \leq \varepsilon_{i_\varepsilon}$; monotonicity of Ψ_ℓ^{-1} then implies $\frac{\Psi_\ell^{-1}(\varepsilon_{i-1})}{\varepsilon_i} = z \leq \frac{\Psi_\ell^{-1}(\varepsilon_{i_\varepsilon})}{\varepsilon_{i_\varepsilon+1}} \leq \frac{\Psi_\ell^{-1}(\Psi_\ell(\varepsilon))}{\Psi_\ell(\varepsilon)/4} = \frac{4\varepsilon}{\Psi_\ell(\varepsilon)}$. Altogether, we have that

$$\frac{p_{i-1}}{\varepsilon_i} \leq \frac{\theta\left(a\varepsilon^\alpha\right)a(4\varepsilon)^\alpha}{\Psi_\ell(\varepsilon)^\alpha\varepsilon_i^{1-\alpha}} \leq \frac{4\theta\left(a\varepsilon^\alpha\right)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}.$$

Therefore,

$$
\begin{aligned}
&\sum_{i=1}^{i_\varepsilon} \log_2(4(1+i)^2/\delta) + 2em_i p_{i-1} \\
&\lesssim i_\varepsilon \mathrm{Log}(i_\varepsilon/\delta) + i_\varepsilon b \left(\frac{\theta\left(a\varepsilon^\alpha\right) a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} \left(d_\ell \mathrm{Log}\left(\frac{1}{b}\left(\frac{\theta\left(a\varepsilon^\alpha\right) a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right. \\
&\qquad\qquad \left. + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right) \\
&\lesssim b \left(\frac{\theta\left(a\varepsilon^\alpha\right) a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} \left(d_\ell \mathrm{Log}\left(\frac{1}{b}\left(\frac{\theta\left(a\varepsilon^\alpha\right) a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{\beta}\right)\right. \\
&\qquad\qquad \left. + \mathrm{Log}\left(\frac{\mathrm{Log}\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)}{\delta}\right)\right) \mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right).
\end{aligned}
$$

Thus, for an appropriate choice of the constant factor c_0, we have

$$n \geq \sum_{i=1}^{i_\varepsilon} \log_2(4(1+i)^2/\delta) + 2em_i p_{i-1}.$$

Let $M \subseteq \{0, \ldots, 2^n\}$ denote the set of values of m obtained during the execution. Let $V_1 = \mathcal{F}$ and $Q_1 = \emptyset$, and for each $m \in M$ with $\log_2(m) \in \mathbb{N}$, let Q_m denote the value of Q upon reaching Step 5 with that value of m, and let V_m denote the value of V upon completion of Step 6′ (i.e., after the update); additionaly, denote $V_m^{\pm} = \{\mathrm{sign}(f) : f \in V_m\}$. Also, for each $m \in M$ with $\log_2(m) \in \mathbb{N}$, define $P_m = \mathcal{P}_{XY}(\cdot|\mathrm{DIS}(V_{m/2}) \times \mathcal{Y})$, the conditional distribution of (X, Y) given $V_{m/2}$ and $X \in \mathrm{DIS}(V_{m/2})$, where $(X, Y) \sim \mathcal{P}_{XY}$. In particular, note that for each $m \in M$ with $\log_2(m) \in \mathbb{N}$, the set Q_m is conditionally i.i.d. P_m given $|Q_m|$ and $V_{m/2}$. Furthermore, the conditional distribution of Y given X is unchanged (almost everywhere) by conditioning on $V_{m/2}$ and $X \in \mathrm{DIS}(V_{m/2})$, so that may take $f^\star_{P_m,\ell} = f^\star_\ell$, which implies $f^\star_{P_m,\ell} \in \mathcal{F}$. Thus, by Lemma 6.4, P_m satisfies Condition 6.3 with the given values of b and β. Applying Lemma 6.6 conditioned on $V_{m/2}$ and $|Q_m|$, together with the law of total probability (to integrate these variables out) and a union bound (over values of m with $\log_2(m) \in \mathbb{N}$), we have that on an event E_δ of probability at least $1 - \sum_{i=1}^{\infty} \frac{\delta}{(1+i)^2} > 1 - 2\delta/3$, every $m \in M$ with $\log_2(m) \in \mathbb{N}$ has

$$\mathrm{R}_\ell(f_\ell^\star; Q_m) - \inf_{g \in \mathcal{F}} \mathrm{R}_\ell(g; Q_m) \leq U_\ell(|Q_m|, \delta_m), \tag{6.13}$$

and $\forall f \in \mathcal{F}$,

$$\begin{aligned} &\mathrm{R}_\ell(f; P_m) - \mathrm{R}_\ell(f_\ell^\star; P_m) \\ &\qquad \leq \max\left\{2(\mathrm{R}_\ell(f; Q_m) - \mathrm{R}_\ell(f_\ell^\star; Q_m)), U_\ell(|Q_m|, \delta_m)\right\}. \end{aligned} \tag{6.14}$$

Therefore, by the definition of V_m in Step 6′, combined with (6.13) and (6.14), on the event E_δ, $f_\ell^\star \in V_m$ and

$$\forall f \in V_m, \mathrm{R}_\ell(f; P_m) - \mathrm{R}_\ell(f_\ell^\star; P_m) \leq 2U_\ell(|Q_m|, \delta_m).$$

Also, for any $f \in V_m$, the function $f_m = f \mathbb{1}_{\mathrm{DIS}(V_{m/2})} + f_\ell^\star \mathbb{1}_{\mathcal{X} \setminus \mathrm{DIS}(V_{m/2})}$ has

$$\begin{aligned} \mathrm{R}_\ell(f_m) - \mathrm{R}_\ell(f_\ell^\star) &= (\mathrm{R}_\ell(f; P_m) - \mathrm{R}_\ell(f_\ell^\star; P_m))\mathcal{P}(\mathrm{DIS}(V_{m/2})) \\ &\leq 2U_\ell(|Q_m|, \delta_m)\mathcal{P}(\mathrm{DIS}(V_{m/2})) \end{aligned}$$

on the event E_δ; furthermore, $V_m \subseteq V_{m/2}$ implies $\mathrm{DIS}(V_m) \subseteq \mathrm{DIS}(V_{m/2})$, and together with the fact that $f_\ell^\star \in V_m$ on E_δ, we have $\mathrm{sign}(f_m) = \mathrm{sign}(f)$; in particular, this implies $\mathrm{sign}(f_m) \in \mathbb{C}$ and that $\mathrm{er}(f) = \mathrm{er}(f_m)$. Therefore, Lemma 6.5 implies that, on the event E_δ, every $m \in M$ with $\log_2(m) \in \mathbb{N}$ has

$$\forall f \in V_m, \mathrm{er}(f) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}\left(2U_\ell(|Q_m|, \delta_m)\mathcal{P}(\mathrm{DIS}(V_{m/2}))\right), \tag{6.15}$$

where Ψ_ℓ^{-1} is the inverse of Ψ_ℓ, which is well-defined in this context since Ψ_ℓ is continuous and strictly increasing on $(0, \infty)$ due to ℓ being classification-calibrated. Since $\Psi_\ell^{-1}(\Psi_\ell(\varepsilon)) = \varepsilon$, we find that it suffices to show that $\exists m \in M$ with $\log_2(m) \in \mathbb{N}$ and $2U_\ell(|Q_m|, \delta_m)\mathcal{P}(\mathrm{DIS}(V_{m/2})) \leq \Psi_\ell(\varepsilon)$.

We now proceed by induction. Suppose, for some $i \in I$, there is an event E'_{i-1} of probability at least $1 - \sum_{j=1}^{i-1} \frac{\delta}{2(1+j)^2}$ such that, on

$E'_{i-1} \cap E_\delta$, $m_{i-1} \in M$, $V^{\pm}_{m_{i-1}} \subseteq \mathbb{C}\left(\Psi_\ell^{-1}(\varepsilon_{i-1})\right)$, and

$$\sum_{\substack{m \leq m_{i-1}: \\ \log_2(m) \in \mathbb{N}}} |Q_m| \leq \sum_{j=1}^{i-1} \log_2(4(1+j)^2/\delta) + 2em_j p_{j-1}.$$

The above claims are trivially satisfied for $i = 1$ (i.e., $m_0 \in M$, $V^{\pm}_{m_0} \subseteq \mathbb{C}(\Psi_\ell^{-1}(\varepsilon_0))$, and $0 \leq 0$), so that we have a base case for this inductive proof. Now fix any $i \in I$ satisfying these claims, and note that the number of labels requested among data points with indices m between $m_{i-1}+1$ and m_i can be expressed as

$$\sum_{j=\log_2(m_{i-1})}^{\log_2(m_i)-1} \sum_{m=2^j+1}^{\min\{2^{j+1}, \max M\}} \mathbb{1}_{\mathrm{DIS}(V_{2^j})}(X_m) \leq \sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}(V_{m_{i-1}})}(X_m).$$

Furthermore, on $E'_{i-1} \cap E_\delta$, this is at most

$$\sum_{m=m_{i-1}+1}^{m_i} \mathbb{1}_{\mathrm{DIS}\left(\mathbb{C}\left(\Psi_\ell^{-1}(\varepsilon_{i-1})\right)\right)}(X_m).$$

This is a sum of $m_i - m_{i-1}$ independent Bernoulli random variables, each with mean p_{i-1}; therefore, by a Chernoff bound, on an event E''_i of probability at least $1 - \frac{\delta}{4(1+i)^2}$, this sum evaluates to at most

$$\log_2\left(4(1+i)^2/\delta\right) + 2em_i p_{i-1}.$$

Since $m_i \leq 2^n$ and

$$\sum_{j=1}^{i} \log_2(4(1+j)^2/\delta) + 2em_j p_{j-1} \leq n,$$

we have that on $E''_i \cap E'_{i-1} \cap E_\delta$, $m_i \in M$ and

$$\sum_{m \leq m_i : \log_2(m) \in \mathbb{N}} |Q_m| \leq \sum_{j=1}^{i} \log_2(4(1+j)^2/\delta) + 2em_j p_{j-1}.$$

Furthermore, by a Chernoff bound (under the conditional distribution given $V_{m_i/2}$) and the law of total probability, there is an event E'''_i of probability at least $1 - \frac{\delta}{4(1+i)^2}$ such that, on $E_i''' \cap E_i'' \cap E_{i-1}' \cap E_\delta$, if $\mathcal{P}\left(\mathrm{DIS}\left(V_{m_i/2}\right)\right) > \left(\frac{16}{m_i}\right) \ln\left(4(1+i)^2/\delta\right)$, then

$$|Q_{m_i}| \geq (m_i/4)\mathcal{P}\left(\mathrm{DIS}\left(V_{m_i/2}\right)\right).$$

Let us extend the domain of $m \mapsto U_\ell(m, \delta_{m_i})$ to $[1, \infty)$, defined by the same formula given in Lemma 6.6; furthermore, note that $m \mapsto U_\ell(m, \delta_{m_i})$ is nonincreasing on $[1, \infty)$, while $m \mapsto U_\ell(m, \delta_{m_i})m$ is nondecreasing on $[1, \infty)$. In particular, this implies that on $E_i''' \cap E_i'' \cap E_{i-1}' \cap E_\delta$, if $\mathcal{P}\left(\mathrm{DIS}\left(V_{m_i/2}\right)\right) > \left(\frac{16}{m_i}\right) \ln\left(4(1+i)^2/\delta\right)$,

$$\begin{aligned}
&2U_\ell\left(|Q_{m_i}|, \delta_{m_i}\right)\mathcal{P}(\mathrm{DIS}(V_{m_i/2})) \\
&\leq 2U_\ell\left((m_i/4)\mathcal{P}(\mathrm{DIS}(V_{m_i/2})), \delta_{m_i}\right)\mathcal{P}(\mathrm{DIS}(V_{m_i/2})) \\
&\leq 2U_\ell\left((m_i/4)\mathcal{P}(\mathrm{DIS}(V_{m_{i-1}})), \delta_{m_i}\right)\mathcal{P}(\mathrm{DIS}(V_{m_{i-1}})) \\
&\leq 2U_\ell\left((m_i/4)p_{i-1}, \delta_{m_i}\right)p_{i-1}.
\end{aligned}$$

For a sufficiently large choice of the constant c'' in the definition of m_i, this last expression above is at most ε_i. Furthermore, if $\mathcal{P}\left(\mathrm{DIS}\left(V_{m_i/2}\right)\right) \leq \left(\frac{16}{m_i}\right) \ln\left(4(1+i)^2/\delta\right)$, then $2U_\ell(|Q_{m_i}|, \delta_{m_i})\mathcal{P}(\mathrm{DIS}(V_{m_i/2})) \leq \left(\frac{32\bar{\ell}}{m_i'}\right) \ln\left(4(1+i)^2/\delta\right) \leq \left(\frac{128\bar{\ell}}{c''}\right)\varepsilon_i$, which is at most ε_i for a sufficiently large choice of the constant c''. Thus, in any case, on the event $E_i''' \cap E_i'' \cap E_{i-1}' \cap E_\delta$, $2U_\ell(|Q_{m_i}|, \delta_{m_i})\mathcal{P}(\mathrm{DIS}(V_{m_i/2})) \leq \varepsilon_i$. Plugging this into (6.15), we find that on $E_i''' \cap E_i'' \cap E_{i-1}' \cap E_\delta$, every $f \in V_{m_i}$ has $\mathrm{er}(f) - \mathrm{er}(f^\star) \leq \Psi_\ell^{-1}(\varepsilon_i)$. Therefore, by a union bound, taking $E_i' = E_i''' \cap E_i'' \cap E_{i-1}'$, we have extended the inductive hypothesis. In particular, by the principle of induction, we have established that there exists an event E_{i_ε}' of probability at least $1 - \sum_{j=1}^{i_\varepsilon} \frac{\delta}{2(1+j)^2} > 1 - \delta/3$ such that, on $E_{i_\varepsilon}' \cap E_\delta$, $m_{i_\varepsilon} \in M$ and $V_{m_{i_\varepsilon}}^{\pm} \subseteq \mathbb{C}\left(\Psi_\ell^{-1}(\varepsilon_{i_\varepsilon})\right) \subseteq \mathbb{C}\left(\Psi_\ell^{-1}\left(\Psi_\ell(\varepsilon)\right)\right) = \mathbb{C}(\varepsilon)$, so that in particular, $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. Also note that, by a union bound, the event $E_{i_\varepsilon}' \cap E_\delta$ has probability greater than $1 - \delta$. □

The bound in Theorem 6.9 represents an improvement over that of Theorem 6.8 (assuming the same values of b and β) in the cases of $r_\ell > 2$, essentially multiplying by a factor proportional to $\left(\theta\left(a\varepsilon^\alpha\right) a\varepsilon^\alpha\right)^{1-\beta}$.

As with the other results above, the final $\mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)$ factor in Theorem 6.9 can be replaced by a constant in many cases, by a careful treatment of the summations in the proof, and the $\mathrm{Log}(\mathrm{Log}(b/\Psi_\ell(\varepsilon)))$ factor can also be replaced by a constant in many cases, after a modification to the definition of δ_m [see Hanneke and Yang, 2012].

6.6 To Optimize or Not to Optimize

We conclude Chapter 6 with a few thoughts on the appropriate uses of surrogate losses in active learning, as compared to passive learning. Presently, the most common approach to label complexity analysis in passive learning with a surrogate loss is to bound the label complexity of producing a function $\hat{f} \in \mathcal{F}$ with $\mathrm{R}_\ell(\hat{f}) - \inf_{g\in\mathcal{F}} \mathrm{R}_\ell(g) \le \gamma$ in terms of an arbitrary value $\gamma > 0$, and then plug in a value of γ sufficiently small to guarantee that any $f \in \mathcal{F}$ with $\mathrm{R}_\ell(f) - \inf_{g\in\mathcal{F}} \mathrm{R}_\ell(g) \le \gamma$ necessarily also has $\mathrm{er}(f) - \mathrm{er}(f^\star) \le \varepsilon$. For instance, we used this approach above to obtain Theorem 6.7, using $\gamma = \Psi_\ell(\varepsilon)$ in that case.

There are now several active learning algorithms in the published literature that are capable of optimizing $\mathrm{R}_\ell(f)$ over $f \in \mathcal{F}$, with provable guarantees on the number of label requests sufficient for them to achieve a desired value of $\mathrm{R}_\ell(f) - \inf_{g\in\mathcal{F}} \mathrm{R}_\ell(g)$; for instance, Beygelzimer, Dasgupta, and Langford [2009] and Koltchinskii [2010] each present variants of the disagreement-based active learning strategy suitable for this task. In light of this, one might think it equally natural to approach the study of the label complexity of active learning with surrogate losses via the same *direct* approach described above for passive learning: that is, to use one of these active learning strategies that has a guarantee on the number of label requests sufficient to produce a function $\hat{f} \in \mathcal{F}$ with $\mathrm{R}_\ell(\hat{f}) - \inf_{g\in\mathcal{F}} \mathrm{R}_\ell(g) \le \gamma$, expressed as a function of γ, and then to plug in a sufficiently small value of γ (e.g., $\Psi_\ell(\varepsilon)$) to guarantee that any $f \in \mathcal{F}$ with $\mathrm{R}_\ell(f) - \inf_{g\in\mathcal{F}} \mathrm{R}_\ell(g) \le \gamma$ necessarily has $\mathrm{er}(f) - \mathrm{er}(f^\star) \le \varepsilon$.

However, interestingly, one can show that this approach *cannot* provide label complexity bounds as strong as Theorems 6.8 and 6.9 above [Hanneke and Yang, 2010, 2012]. Specifically, one can construct very natural scenarios where the label complexity of $\mathrm{ERM}_\ell(\mathcal{F},\cdot)$ is $\Theta(1/\varepsilon)$, and where the label complexity bounds above are $O(\log(1/\varepsilon)\log\log(1/\varepsilon))$, indicating strong improvements over passive learning, but where it is *not* possible with fewer than $\Theta(1/\varepsilon)$ label requests to produce a function $\hat{f}$ with a guarantee that $\mathrm{R}_\ell(\hat{f}) - \inf_{g\in\mathcal{F}}\mathrm{R}_\ell(g) \leq \gamma$ (with reasonably high probability), for a value γ sufficiently small to guarantee *every* $f \in \mathcal{F}$ with $\mathrm{R}_\ell(f) - \inf_{g\in\mathcal{F}}\mathrm{R}_\ell(g) \leq \gamma$ has $\mathrm{er}(f) - \mathrm{er}(f^\star) \leq \varepsilon$. Thus, the key insight enabling us to obtain these strong improvements in label complexity over passive learning is that we are only interested in ℓ as a computational tool, which helps us to optimize $\mathrm{er}(h)$ over $h \in \mathbb{C}$, and as long as this end is attained, we have no real interest in further optimizing the value of $\mathrm{R}_\ell(f)$. In fact, $\mathrm{RobustCAL}_\delta^\ell$ typically *does not optimize* $\mathrm{R}_\ell(f)$ over $f \in \mathcal{F}$ (even in the limit of $n \to \infty$). In this sense, the appropriate use of surrogate losses in active learning seems quite different from that in passive learning (though Hanneke and Yang, 2012, find that these insights sometimes have implications for passive learning as well).

7

Bounding the Disagreement Coefficient

The results of the previous sections expressed bounds on the label complexity of active learning in terms of the disagreement coefficient. As such, we are clearly interested in obtaining bounds on the disagreement coefficient itself, for various learning problems of interest, since such bounds would compose with the above to imply results on the label complexity. The disagreement coefficient $\theta_h(\varepsilon)$ has been studied and bounded under various conditions on $\mathbb{C}$, $\mathcal{P}$, and h. In this section, we survey these findings, along with a few previously-unpublished results.

First, in Section 7.1, we describe a few very basic properties and inequalities that the disagreement coefficient always satisfies. This is followed in Section 7.2 by a discussion of the asymptotic dependence on ε in $\theta_h(\varepsilon)$, including techniques that can help to simplify the process of characterizing this dependence. In Section 7.3, we discuss a kind of coarse analysis of the disagreement coefficient, in particular stating conditions that are sufficient to guarantee $\theta_h(\varepsilon) = o(1/\varepsilon)$, and stronger conditions sufficient to guarantee $\theta_h(\varepsilon) = O(1)$; for simplicity, we focus primarily on linear separators in that section, and find that any $\mathcal{P}$ that has a density function p guarantees $\theta_h(\varepsilon) = o(1/\varepsilon)$; furthermore, if p is bounded and has bounded support, and the separating hyperplane for h passes through a continuity point of p in the support of p, then $\theta_h(\varepsilon) = O(1)$. We also discuss generalizations of these results to a larger family of hypothesis classes. In Section 7.4, we briefly survey the more-detailed known results for a few different hypothesis classes; for context, that section also includes descriptions of other known results on the label complexity of active learning for these hypothesis classes. Section 7.5 describes a general technique for constructing hypothesis classes and distributions such that $\theta_h(\varepsilon)$ realizes any function of ε bounded by $1/\varepsilon$. Finally, Section 7.6 briefly describes some of the more-interesting behaviors that have been noted about the disagreement coefficient, when the class $\mathbb{C}$ is countable. Sections 7.5 and 7.6 are more technical than the others, and can be skipped by the casual reader without significant loss of continuity.

7.1 Basic Properties

We begin by going through a few basic properties of the disagreement coefficient, which always hold. Throughout this section and the next, we fix an arbitrary classifier h (not necessarily in $\mathbb{C}$), and discuss properties of the $\theta_h(\cdot)$ function.

Theorem 7.1. $x \mapsto \theta_h(x)x$ is nondecreasing on $[0, \infty)$, while $x \mapsto \theta_h(x)$ is nonincreasing on $[0, \infty)$.

Proof. That $x \mapsto \theta_h(x)$ is nonincreasing is clear from the definition, due to the supremum. To prove the first claim, fix any values x and y with $0 \leq x < y < \infty$. We have

$$\begin{aligned}
\theta_h(x)x &= x \vee \sup_{r>x} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} x \\
&= x \vee \max\left\{ \sup_{x<r\leq y} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} x, \sup_{r>y} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} x \right\} \\
&\leq y \vee \max\left\{ \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,y))), \sup_{r>y} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} y \right\} \\
&= y \vee \sup_{r\geq y} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} y.
\end{aligned}$$

Since $r \mapsto \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))$ is nondecreasing, this last expression equals $\theta_h(y)y$. $\square$

Corollary 7.2. Let $\varepsilon \in (0,\infty)$ and $c \in (1,\infty)$. Then $\theta_h(\varepsilon/c) \leq c\theta_h(\varepsilon)$ and $\theta_h(\varepsilon)/c \leq \theta_h(c\varepsilon)$.

Proof. Since $\varepsilon/c < \varepsilon$, Theorem 7.1 implies $\theta_h(\varepsilon/c)(\varepsilon/c) \leq \theta_h(\varepsilon)\varepsilon$. Therefore, $\theta_h(\varepsilon/c) = (\theta_h(\varepsilon/c)(\varepsilon/c))(c/\varepsilon) \leq (\theta_h(\varepsilon)\varepsilon)(c/\varepsilon) = c\theta_h(\varepsilon)$. That $\theta_h(\varepsilon)/c \leq \theta_h(c\varepsilon)$ follows by substituting $\varepsilon \leftarrow c\varepsilon$ in the above. $\square$

Theorem 7.3. $x \mapsto \theta_h(x)$ is continuous on $(0,\infty)$. Furthermore, $\theta_h(0) = \lim_{\varepsilon\to 0} \theta_h(\varepsilon)$.

Proof. For any $\varepsilon_0 \in (0,\infty)$ and $\delta \in (-1,1)$, Theorem 7.1 and Corollary 7.2 imply

$$\theta_h(\varepsilon_0)/(1+|\delta|) \leq \theta_h((1+\delta)\varepsilon_0) \leq \theta_h(\varepsilon_0)/(1-|\delta|),$$

so that $\lim_{\varepsilon\to\varepsilon_0} \theta_h(\varepsilon) = \lim_{|\delta|\to 0} \theta_h((1+\delta)\varepsilon_0) = \theta_h(\varepsilon_0)$.

That $\theta_h(0) = \lim_{\varepsilon\to 0} \theta_h(\varepsilon)$ follows from continuity of the supremum from below: that is, for any nondecreasing sequence $\{A_i\}_{i=1}^{\infty}$ of nonempty subsets of $\mathbb{R}$, $\sup\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty}(\sup A_n)$; with $A_n = \{1 \vee \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))/r : r > 1/n\}$, this implies the claim. $\square$

In addition to this continuity of $\theta_h(\varepsilon)$ in ε, we also have continuity in h (for $\varepsilon > 0$), as implied by the following result.

Theorem 7.4. Let $\{h_i\}_{i=1}^{\infty}$ be any sequence of classifiers (not necessarily in $\mathbb{C}$) with $\lim_{i\to\infty} \mathcal{P}(x : h_i(x) \neq h(x)) = 0$. Then $\forall \varepsilon > 0$, $\lim_{i\to\infty} \theta_{h_i}(\varepsilon) = \theta_h(\varepsilon)$.

Proof. Since $\lim_{i\to\infty} \mathcal{P}(x : h_i(x) \neq h(x)) = 0$, for any $\gamma > 0$, $\exists i_\gamma \in \mathbb{N}$ s.t. $\forall i \geq i_\gamma$, $\mathcal{P}(x : h_i(x) \neq h(x)) \leq \gamma$. In particular, this implies that for every $i \geq i_\gamma$, $\forall r > 0$, $\mathrm{B}(h_i, r+\gamma) \supseteq \mathrm{B}(h,r)$ and $\mathrm{B}(h, r+\gamma) \supseteq \mathrm{B}(h_i, r)$.

Therefore, for any $\varepsilon > 0$,

$$\begin{aligned}
\theta_{h_i}(\varepsilon) &= 1 \vee \sup_{r>\varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_i, r)))}{r} \leq 1 \vee \sup_{r>\varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r+\gamma)))}{r} \\
&\leq \frac{\varepsilon+\gamma}{\varepsilon}\left(1 \vee \sup_{r>\varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r+\gamma)))}{r+\gamma}\right) \\
&= \frac{\varepsilon+\gamma}{\varepsilon}\theta_h(\varepsilon+\gamma) \leq \frac{\varepsilon+\gamma}{\varepsilon}\theta_h(\varepsilon).
\end{aligned}$$

Similarly, $\theta_h(\varepsilon) \leq \frac{\varepsilon+\gamma}{\varepsilon}\theta_{h_i}(\varepsilon)$. In particular, since $i_\gamma < \infty$ for every $\gamma > 0$, this implies that $\forall \gamma > 0$, $\limsup\limits_{i\to\infty} \theta_{h_i}(\varepsilon) \leq \frac{\varepsilon+\gamma}{\varepsilon}\theta_h(\varepsilon)$ and $\frac{\varepsilon}{\varepsilon+\gamma}\theta_h(\varepsilon) \leq \liminf\limits_{i\to\infty} \theta_{h_i}(\varepsilon)$. Taking the limit as $\gamma \to 0$, we have $\limsup\limits_{i\to\infty} \theta_{h_i}(\varepsilon) \leq \theta_h(\varepsilon) \leq \liminf\limits_{i\to\infty} \theta_{h_i}(\varepsilon)$, so that $\lim\limits_{i\to\infty} \theta_{h_i}(\varepsilon)$ exists and equals $\theta_h(\varepsilon)$. $\square$

Below, we discuss bounds on $\theta_h(\varepsilon)$ holding under conditions on h, $\mathbb{C}$, and $\mathcal{P}$. However, there are also some very simple bounds on $\theta_h(\varepsilon)$, which always hold. For instance, since probabilities are at most 1, we clearly always have $\theta_h(\varepsilon) \leq 1/\varepsilon$. For finite classes $\mathbb{C}$, we also have the following basic result.

Theorem 7.5. $\theta_h(0) \leq |\mathbb{C}|$.

Proof. Since $\forall r > 0$,

$$\mathrm{DIS}(\mathrm{B}(h,r)) \subseteq \mathrm{DIS}(\mathrm{B}(h,r) \cup \{h\}) = \bigcup_{g \in \mathrm{B}(h,r)} \mathrm{DIS}(\{h,g\}),$$

a union bound implies

$$\begin{aligned}
\theta_h(0) &= \sup_{r>0} \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(h,r\right)\right)\right)}{r} \leq \sup_{r>0} \frac{\sum_{g\in\mathrm{B}(h,r)} \mathcal{P}(x : g(x) \neq h(x))}{r} \\
&\leq \sup_{r>0} \frac{\sum_{g\in\mathrm{B}(h,r)} r}{r} = \sup_{r>0} |\mathrm{B}(h,r)| = |\mathbb{C}|.
\end{aligned}$$

$\square$

Suppose we are able to bound the value of $\theta_h(\varepsilon)$ with respect to some particular hypothesis classes $\mathbb{C}$ and under certain distributions $\mathcal{P}$. There are then several properties of the disagreement coefficient which immediately allow us to generalize this to results for a whole family of classes $\mathbb{C}$ and distributions $\mathcal{P}$. Specifically, we have the following properties, from Hanneke [2011].

Theorem 7.6. Let $\lambda \in (0,1)$, and suppose $\mathcal{P}$ and $\mathcal{P}'$ are distributions over $\mathcal{X}$ such that $\lambda \mathcal{P}' \leq \mathcal{P} \leq (1/\lambda)\mathcal{P}'$. For all $\varepsilon > 0$, let $\theta_h(\varepsilon)$ and $\theta'_h(\varepsilon)$ denote the disagreement coefficients of h with respect to $\mathbb{C}$ under $\mathcal{P}$ and $\mathcal{P}'$, respectively. Then $\forall \varepsilon > 0$,

$$\theta'_h(\varepsilon\lambda)\lambda^2 \leq \theta_h(\varepsilon) \leq \theta'_h(\varepsilon/\lambda)/\lambda^2.$$

Proof. We prove the first inequality; the second inequality follows from the first, since we also have $\lambda\mathcal{P} \leq \mathcal{P}' \leq (1/\lambda)\mathcal{P}$, so that reversing the roles of the two distributions and dividing the ε argument by λ yields the second inequality.

For any $g \in \mathbb{C}$, $\lambda\mathcal{P}(x : h(x) \neq g(x)) \leq \mathcal{P}'(x : h(x) \neq g(x))$. Thus, $\forall r > 0$, $\mathrm{B}_{\mathcal{P}'}(h, r\lambda) \subseteq \mathrm{B}_{\mathcal{P}}(h, r)$, which implies

$$\lambda\mathcal{P}'\left(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r\lambda))\right) \leq \mathcal{P}\left(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r\lambda))\right) \leq \mathcal{P}\left(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))\right).$$

We therefore have

$$\begin{aligned}\lambda^2 \sup_{r > \varepsilon\lambda} \frac{\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r)))}{r} &= \sup_{r > \varepsilon} \frac{\lambda\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r\lambda)))}{r} \\ &\leq \sup_{r > \varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r)))}{r}.\end{aligned}$$

The first inequality in the theorem immediately follows from this and the definition of the disagreement coefficient. □

Theorem 7.7. Suppose there exist $\lambda \in (0,1)$ and distributions $\mathcal{P}'$ and $\mathcal{P}''$ over $\mathcal{X}$ such that $\mathcal{P} = \lambda\mathcal{P}' + (1-\lambda)\mathcal{P}''$. For $\varepsilon > 0$, let $\theta_h(\varepsilon)$, $\theta'_h(\varepsilon)$, and $\theta''_h(\varepsilon)$ denote the disagreement coefficients of h with respect to $\mathbb{C}$ under $\mathcal{P}$, $\mathcal{P}'$, and $\mathcal{P}''$, respectively. Then $\forall \varepsilon > 0$,

$$\theta_h(\varepsilon) \leq \theta'_h(\varepsilon/\lambda) + \theta''_h(\varepsilon/(1-\lambda)).$$

Proof. For any $r > 0$,

$$\begin{aligned}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r))) &= \lambda\mathcal{P}'(\mathrm{DIS}(\mathrm{B}(h, r))) + (1-\lambda)\mathcal{P}''(\mathrm{DIS}(\mathrm{B}(h, r))) \\ &\leq \lambda\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r/\lambda))) + (1-\lambda)\mathcal{P}''(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}''}(h, r/(1-\lambda)))).\end{aligned}$$

Thus,

$$\begin{aligned}
&\sup_{r>\varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))}{r} \\
&\leq \sup_{r>\varepsilon} \frac{\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r/\lambda)))}{r/\lambda} + \sup_{r>\varepsilon} \frac{\mathcal{P}''(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}''}(h,r/(1-\lambda))))}{r/(1-\lambda)} \\
&\leq \theta'_h(\varepsilon/\lambda) + \theta''_h(\varepsilon/(1-\lambda)).
\end{aligned}$$

The result now follows from the definition of the disagreement coefficient. $\square$

Theorem 7.8. Let $\mathbb{C}'$ and $\mathbb{C}''$ be sets of classifiers such that $\mathbb{C} = \mathbb{C}' \cup \mathbb{C}''$, and let $\mathcal{P}$ be a distribution over $\mathcal{X}$. For all $\varepsilon > 0$, let $\theta_h(\varepsilon)$, $\theta'_h(\varepsilon)$, and $\theta''_h(\varepsilon)$ denote the disagreement coefficients of h with respect to $\mathbb{C}$, $\mathbb{C}'$, and $\mathbb{C}''$, respectively, under $\mathcal{P}$. Then $\forall \varepsilon > 0$,

$$\max\left\{\theta'_h(\varepsilon), \theta''_h(\varepsilon)\right\} \leq \theta_h(\varepsilon) \leq \theta'_h(\varepsilon) + \theta''_h(\varepsilon) + 2.$$

Furthermore, if $h \in \mathbb{C}$, then $\theta_h(\varepsilon) \leq \theta'_h(\varepsilon) + \theta''_h(\varepsilon) + 1$, and if $h \in \mathbb{C}' \cap \mathbb{C}''$, then $\theta_h(\varepsilon) \leq \theta'_h(\varepsilon) + \theta''_h(\varepsilon)$.

Proof. The first inequality is clear from the definition of the disagreement coefficient (due to monotonicity of $\mathcal{H} \mapsto \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{H}}(h,r)))$). To prove the second inequality, note that $\forall r > 0$,

$$\begin{aligned}
&\mathrm{DIS}(\mathrm{B}_{\mathbb{C}}(h,r)) \\
&= \mathrm{DIS}(\mathrm{B}_{\mathbb{C}'}(h,r)) \cup \mathrm{DIS}(\mathrm{B}_{\mathbb{C}''}(h,r)) \cup \bigcap_{\substack{f \in \mathrm{B}_{\mathbb{C}'}(h,r), \\ g \in \mathrm{B}_{\mathbb{C}''}(h,r)}} \mathrm{DIS}(\{f,g\}).
\end{aligned}$$

Let $\Delta = \inf_{f \in \mathrm{B}_{\mathbb{C}'}(h,r)} \inf_{g \in \mathrm{B}_{\mathbb{C}''}(h,r)} \mathcal{P}(x : f(x) \neq g(x))$ if $\mathrm{B}_{\mathbb{C}'}(h,r) \neq \emptyset$ and $\mathrm{B}_{\mathbb{C}''}(h,r) \neq \emptyset$, and let $\Delta = 0$ otherwise. By a union bound,

$$\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathbb{C}}(h,r))) \leq \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathbb{C}'}(h,r))) + \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathbb{C}'}(h,r))) + \Delta.$$

Noting that any $f \in \mathrm{B}_{\mathbb{C}'}(h,r)$ and $g \in \mathrm{B}_{\mathbb{C}''}(h,r)$ have $\mathcal{P}(x : f(x) \neq g(x)) \leq \mathcal{P}(x : f(x) \neq h(x)) + \mathcal{P}(x : g(x) \neq h(x)) \leq 2r$, we have $\Delta \leq 2r$. The inequality now follows from the definition of the disagreement coefficient by dividing both sides of this inequality by r and taking the supremum over $r > \varepsilon$. The two stronger results follow from this as well. Specifically, if $h \in \mathbb{C}$, then we can bound Δ by taking one of f or g to be h, in which case the other being in $\mathrm{B}(h, r)$ entails $\Delta \leq r$. Furthermore, if $h \in \mathbb{C}' \cap \mathbb{C}''$, then $\mathrm{B}_{\mathbb{C}'}(h, r) \cap \mathrm{B}_{\mathbb{C}''}(h, r) \neq \emptyset$, so that $\Delta = 0$. $\square$

7.2 Asymptotic Behavior

There are several general analyses that involve the asymptotic dependence of $\theta_h(\varepsilon)$ on ε [e.g., Balcan, Hanneke, and Vaughan, 2010, Friedman, 2009]. Combined with the results of Chapter 5, such analyses can be used to characterize the asymptotic dependence on ε in the label complexity of disagreement-based active learning algorithms. In the context of this type of analysis, it is often easier to study the asymptotic behavior of $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))$ directly, rather than that of $\theta_h(\varepsilon)$. Fortunately, the following lemma allows us to do so without loss.

Lemma 7.9. For any nonincreasing $g : (0, 1) \to [1, \infty)$,

$$\theta_h(\varepsilon) = O(g(\varepsilon)) \text{ iff } \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))}{\varepsilon} = O(g(\varepsilon)), \tag{7.1}$$

and if $g(\varepsilon) = \omega(1)$,

$$\theta_h(\varepsilon) = o(g(\varepsilon)) \text{ iff } \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))}{\varepsilon} = o(g(\varepsilon)). \tag{7.2}$$

Proof. Fix a function g as described. We clearly have

$$\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))}{\varepsilon} \leq \sup_{r > \varepsilon} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r)))}{r} \leq \theta_h(\varepsilon),$$

so that the "only if" half of both claims is obvious.

To prove the "if" half, suppose $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))/\varepsilon = O(g(\varepsilon))$; then there exist constants $\varepsilon_0 \in (0, 1)$ and $c \in [1, \infty)$ such that $\forall r \in (0, \varepsilon_0]$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r)))/r \leq c g(r)$. Thus, $\forall r \in (0, 1)$, we generally have $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r)))/r \leq \theta_h(\varepsilon_0) \vee c g(r) \leq (1/\varepsilon_0) \vee c g(r)$. Therefore, $\forall \varepsilon \in (0, 1)$,

$$\theta_h(\varepsilon) \leq \sup_{r \in (\varepsilon, 1)} ((1/\varepsilon_0) \vee c g(r)) \leq (1/\varepsilon_0) \vee c g(\varepsilon) = O(g(\varepsilon)).$$

Likewise, if $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))/\varepsilon = o(g(\varepsilon))$, then $\forall \delta \in (0, 1)$, $\exists \varepsilon_\delta \in (0, 1)$ such that $\forall r \in (0, \varepsilon_\delta]$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r)))/(r g(r)) < \delta$; furthermore, if we also have $g(\varepsilon) = \omega(1)$, then $\exists r_\delta \in (0,1)$ such that $g(r_\delta) \geq 1/(\delta\varepsilon_\delta)$. Thus, $\forall r \in (0,1)$, we generally have $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,r)))/r \leq \theta_h(\varepsilon_\delta) \vee \delta g(r) \leq (1/\varepsilon_\delta) \vee \delta g(r) \leq \delta g(r_\delta) \vee \delta g(r) = \delta g(r_\delta \wedge r)$. Therefore, $\forall \delta \in (0,1)$, every $\varepsilon \in (0, r_\delta)$ satisfies

$$\theta_h(\varepsilon) \leq \sup_{r \in (\varepsilon,1)} \delta g(r_\delta \wedge r) \leq \delta g(\varepsilon),$$

so that $\lim_{\varepsilon \to 0} \theta_h(\varepsilon)/g(\varepsilon) = 0$. □

In particular, in light of the discussion in Chapter 5, we are clearly interested in scenarios in which $\theta_h(\varepsilon) = O(1)$ (equivalently, $\theta_h(0) < \infty$), since these provide the strongest positive results in Chapter 5. Furthermore, these scenarios allow us the conceptual simplicity of regarding $\theta_h(\varepsilon)$ as a *constant*; more precisely, in this case, we may replace $\theta_h(\varepsilon)$ with its finite constant upper bound $\theta_h(0)$. The following application of Lemma 7.9 can be used to simplify the process of proving $\theta_h(\varepsilon) = O(1)$.

Corollary 7.10. $\theta_h(\varepsilon) = O(1)$ iff $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon))) = O(\varepsilon)$.

Proof. Taking $g(\varepsilon) = 1$ in Lemma 7.9, (7.1) implies the equivalence. □

Beyond this strong $\theta_h(\varepsilon) = O(1)$ behavior, we are also interested in scenarios in which, more generally, $\theta_h(\varepsilon) = o(1/\varepsilon)$, since these also yield positive results about the label complexity advantages of CAL and RobustCAL over their passive learning counterparts (see the discussion near the end of Section 5.1), albeit a somewhat weaker type of improvement than implied by $\theta_h(\varepsilon) = O(1)$. To study this weaker type of improvement, we first introduce a notion of the limiting region of disagreement, referred to as the *disagreement core* by Hanneke [2012]. This region will be a focus of analysis in several contexts below.

Definition 7.11. For any classifier h and set of classifiers $\mathcal{H}$, define the *disagreement core* of h with respect to $\mathcal{H}$ under $\mathcal{P}$ as

$$\partial_{\mathcal{H}} h = \lim_{r \to 0} \mathrm{DIS}(\mathrm{B}_{\mathcal{H}}(h,r)).$$

For $\mathcal{H} = \mathbb{C}$, abbreviate this as $\partial h = \partial_{\mathbb{C}} h$.

The following lemma shows a fundamental relationship between the probability mass in the disagreement core and the condition of $\theta_h(\varepsilon) = o(1/\varepsilon)$, thus formally relating the disagreement core and the disagreement coefficient.

Lemma 7.12. $\theta_h(\varepsilon) = o(1/\varepsilon)$ iff $\mathcal{P}(\partial h) = 0$.

Proof. The continuity of measures implies $\lim_{\varepsilon \to 0} (\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon)))/\varepsilon)/(1/\varepsilon) = \lim_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon))) = \mathcal{P}\left(\lim_{\varepsilon \to 0} \mathrm{DIS}(\mathrm{B}(h,\varepsilon))\right) = \mathcal{P}(\partial h)$. Thus, $\mathcal{P}(\partial h) = 0$ if and only if $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon)))/\varepsilon = o(1/\varepsilon)$. Furthermore, by taking $g(\varepsilon) = 1/\varepsilon$ in Lemma 7.9, (7.2) implies $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon)))/\varepsilon = o(1/\varepsilon)$ if and only if $\theta_h(\varepsilon) = o(1/\varepsilon)$. $\square$

Since we are interested in scenarios in which $\theta_h(\varepsilon) = o(1/\varepsilon)$, it is worth mentioning one particularly general observation: namely, that every *discrete* distribution $\mathcal{P}$ has $\theta_h(\varepsilon) = o(1/\varepsilon)$, regardless of h or $\mathbb{C}$. This is formally summarized in the following results.

Lemma 7.13. $\forall r \in [0,1]$, $\mathrm{DIS}(\mathrm{B}(h,r)) \subseteq \{x : \mathcal{P}(\{x\}) \le r\}$.

Proof. For any $x' \in \mathcal{X}$ with $\mathcal{P}(\{x'\}) > r$, any $g \in \mathbb{C}$ with $g(x') \neq h(x')$ has $\mathcal{P}(x : g(x) \neq h(x)) \ge \mathcal{P}(\{x'\}) > r$, so that $g \notin \mathrm{B}(h,r)$; hence $x' \notin \mathrm{DIS}(\mathrm{B}(h,r))$. $\square$

Theorem 7.14. If $\exists \{x_i\}_{i \in \mathbb{N}}$ in $\mathcal{X}$ such that $\mathcal{P}(\{x_i : i \in \mathbb{N}\}) = 1$, then $\theta_h(\varepsilon) = o(1/\varepsilon)$. In particular, this is true for all $\mathcal{P}$ if $\mathcal{X}$ is countable.

Proof. By Lemma 7.13,

$$\partial h = \lim_{\varepsilon \to 0} \mathrm{DIS}(\mathrm{B}(h,\varepsilon)) \subseteq \lim_{\varepsilon \to 0} \{x : \mathcal{P}(\{x\}) \le \varepsilon\} = \{x : \mathcal{P}(\{x\}) = 0\}. \tag{7.3}$$

Since $\mathcal{P}(\{x_i : i \in \mathbb{N}\}) = 1$, we have $\mathcal{P}(\partial h) = \mathcal{P}(\{x_i : i \in \mathbb{N}\} \cap \partial h)$; combined with (7.3), and monotonicity and additivity of measures, this implies

$$\mathcal{P}(\partial h) \le \mathcal{P}(\{x_i : i \in \mathbb{N}, \mathcal{P}(\{x_i\}) = 0\}) = \sum_{i \in \mathbb{N} : \mathcal{P}(\{x_i\}) = 0} \mathcal{P}(\{x_i\}) = 0.$$

The conclusion now follows from Lemma 7.12. $\square$

7.3 Coarse Analyses under General Conditions

In addition to the above general properties, one can formulate concrete general sufficient conditions, under which the disagreement coefficient can be bounded in interesting ways. This section describes two types of such results: namely, conditions sufficient for $\theta_h(\varepsilon) = o(1/\varepsilon)$ and stronger conditions sufficient for $\theta_h(\varepsilon) = O(1)$. To keep the presentation simple, explicit proofs of these results are given only for the special case of linear separators. In this case, we find $\theta_h(\varepsilon) = o(1/\varepsilon)$ is guaranteed as long as $\mathcal{P}$ has a density; the stronger $\theta_h(\varepsilon) = O(1)$ guarantee is obtained as long as the density is bounded, has bounded support, and the separating hyperplane of h passes through the support at a continuity point of the density. Although these results are only formally proven for linear separators here, the intuition and formal arguments generalize in various natural ways to other hypothesis classes as well (e.g., those with smoothly-parametrized decision boundaries, which are therefore locally approximately linear). In particular, we also describe a natural generalization of the latter result to more general families of hypothesis classes, due to Friedman [2009], but refer the interested reader to the original source for the proof of this more general result.

7.3.1 General Analysis: Linear Separators

This subsection is concerned with the asymptotic behaviors of the disagreement coefficients of linear separators. In this context, we will make use of the following notational conventions. Throughout, we fix an arbitrary $k \in \mathbb{N}$, and suppose $\mathcal{X} = \mathbb{R}^k$. Let $Z_k = \{(z_1, \ldots, z_{k+1}) \in \mathbb{R}^{k+1} : \|(z_1, \ldots, z_k)\| > 0\}$. For any $z = (z_1, \ldots, z_{k+1}) \in Z_k$, let h_z denote the k-dimensional linear separator specified by the weight vector $(z_1, \ldots, z_k)$ and bias z_{k+1}: that is, for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, $h_z(x) = +1$ if $z_{k+1} + \sum_{i=1}^k z_i x_i \geq 0$, and $h_z(x) = -1$ otherwise. Furthermore, let ∂z denote the separating hyperplane associated with h_z: that is, $\partial z = \{x \in \mathbb{R}^k : z_{k+1} + \sum_{i=1}^k z_i x_i = 0\}$. Finally, for any set $S \subseteq \mathbb{R}^k$, let $\operatorname{diam}(S) = \sup_{x,y \in S} \|x - y\|$ denote the diameter of S in the Euclidean metric.

We denote by λ^k the Lebesgue measure on $\mathbb{R}^k$. Recall that we say a probability measure P over $\mathbb{R}^k$ has a *density* $p : \mathbb{R}^k \to [0, \infty]$ (with respect to λ^k) if the function p is measurable, and for all measurable sets $A \subseteq \mathbb{R}^k$, $P(A) = \int \mathbb{1}_A(x)p(x)\lambda^k(\mathrm{d}x)$. In particular, the Radon-Nikodym theorem states that P has a density if and only if every measurable set $A \subseteq \mathbb{R}^k$ satisfies $\lambda^k(A) = 0 \Rightarrow P(A) = 0$. In the results below, we will be particularly interested in distributions $\mathcal{P}$ that have a density p. Interestingly, we find that the mere existence of a density is already sufficient to guarantee $\theta_h(\varepsilon) = o(1/\varepsilon)$. Furthermore, with a few additional conditions on p and h, we can obtain the stronger guarantee that $\theta_h(\varepsilon) = O(1)$. We first give formal statements of both of these results, along with rough outlines of their proofs, before delving into the details of their respective proofs.

We begin with a result indicating $\theta_h(\varepsilon) = o(1/\varepsilon)$ holds whenever $\mathcal{P}$ has a density.

Theorem 7.15. If $\mathbb{C}$ is the class of k-dimensional linear separators, and $\mathcal{P}$ has a density (with respect to λ^k), then $\forall h \in \mathbb{C}$, $\theta_h(\varepsilon) = o(1/\varepsilon)$.

The basic idea of the proof is that, if we let S^+ and S^- denote the smallest convex sets labeled by h as $+1$ and -1, respectively, such that $\mathcal{P}(S^+ \cup S^-) = 1$, then for any point x in either of these regions, any separator that disagrees with h on that point must also disagree with h on some set of points between x and the boundary of that region. By showing that these disagreement sets have probability bounded away from zero, we establish that $x \notin \partial h$. Since this is true of every $x \in S^+ \cup S^-$, and $\mathcal{P}(S^+ \cup S^-) = 1$, we have $\mathcal{P}(\partial h) = 0$, and hence Lemma 7.12 implies $\theta_h(\varepsilon) = o(1/\varepsilon)$. This argument is formalized below in Section 7.3.2. Interestingly, one byproduct of this argument is that, if the separating hyperplane ∂z of h_z intersects the interior of the support of the density, then $\partial h_z = \partial z$: that is, the disagreement core is precisely the set of points on the separating hyperplane.

The above theorem establishes a general sufficient condition for $\theta_h(\varepsilon) = o(1/\varepsilon)$ for the class of linear separators: namely, the existence of a density function. We should note that this is certainly not a necessary condition, and can be substantially relaxed; for instance, we can clearly allow point-masses in addition, by a combination of Theorem 7.7, Theorem 7.14, and Theorem 7.15.

As mentioned, results such as Theorem 7.15 proving $\theta_h(\varepsilon) = o(1/\varepsilon)$ are certainly interesting, since they indicate that disagreement-based active learning methods offer some benefits compared with their passive learning counterparts under these conditions. However, we are often interested in a more detailed description of the *magnitudes* of these benefits, beyond the basic $o(1/\varepsilon)$ claim. We are especially interested in determining when $\theta_h(\varepsilon)$ is *bounded*; as we have seen, these scenarios are particularly important, as they provide the strongest guarantees when combined with the results of Chapter 5. As we discuss below, the mere existence of a density for $\mathcal{P}$ is not sufficient to guarantee $\theta_h(\varepsilon) = O(1)$ for all linear separators h. However, under slightly stronger conditions on $\mathcal{P}$ and h, we can obtain such a result.

The basic idea is that, if the decision boundary of h_z passes through the support of the density p at a continuity point of p, then for any z' close to z, $h_{z'}$ will disagree with h_z in some small region of near-uniform density, and the probability mass they disagree upon in this region will be roughly proportional to either the angle between $(z_1, \ldots, z_k)$ and $(z'_1, \ldots, z'_k)$, or the difference $|z_{k+1} - z'_{k+1}|$. So $\mathrm{B}(h_z, \varepsilon)$ contains only those $h_{z'}$ with this angle and bias difference bounded by $O(\varepsilon)$. If we define

$$\mathrm{supp}(p) = \{x : p(x) > 0\},$$

the *support* of p, then under the condition that the support of p is bounded, a little trigonometry reveals that even the most extreme separators satisfying these angle and bias constraints will not disagree with h_z on any points in the support having distance further that $O(\varepsilon)$ from ∂z; thus, $\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon)) \cap \mathrm{supp}(p)$ is contained in a slab around ∂z of width $O(\varepsilon)$. If p is also bounded, then the probability mass contained in this slab is at most $O(\varepsilon)$, which, by Corollary 7.10, suffices to prove the claim. These conditions are summarized in the following theorem; the formal proof is provided in Section 7.3.3.

Theorem 7.16. If $\mathbb{C}$ is the class of k-dimensional linear separators, and $\mathcal{P}$ has a bounded density p (with respect to λ^k) with $\mathrm{diam}(\mathrm{supp}(p)) < \infty$, then any $z \in Z_k$ such that $\partial z \cap \mathrm{supp}(p)$ contains a continuity point of p has $\theta_{h_z}(\varepsilon) = O(1)$.

For instance, taking $\mathcal{P}$ as a uniform distribution in a compact full-dimensional region, such as a ball or hyper-rectangle, would satisfy the conditions on p in the theorem, though the result also holds under more-interesting distributions as well. The condition that $\partial z \cap \mathrm{supp}(p)$ contains a continuity point of p can be weakened without significantly altering the proof; we merely require that p be bounded away from zero in a neighborhood of some point in ∂z. This latter condition can be further weakened in some cases, but it cannot be entirely removed; for instance, when $\mathcal{P}$ is uniform in a ball in $\mathcal{X} = \mathbb{R}^k$, for $k \geq 2$, any h_z for which ∂z does not intersect the interior of the ball will have an unbounded disagreement coefficient (though still $o(1/\varepsilon)$, due to Theorem 7.15). Likewise, the condition of bounded support, though not always necessary, cannot be entirely removed; in fact, if we allow unbounded support, one can essentially implement the construction of arbitrary $\theta_h(\varepsilon) = o(1/\varepsilon)$ functions presented below in Section 7.5, using linear separators h_z, while maintaining the other conditions on $\mathcal{P}$ and z stated in Theorem 7.16.

7.3.2 Proof of Theorem 7.15

Here, we present the formal proof of Theorem 7.15. We continue the notational conventions introduced above. In addition, to make the argument from above formal, we will make use of the following basic definition and lemmas, which extend a result of Witsenhausen [1968] for distributions over $\mathbb{R}$ to the multidimensional setting. For any probability measure P over $\mathbb{R}^k$, define the set

$$S_P = \{x \in \mathbb{R}^k : \forall z \in Z_k, \text{ if } x \in \partial z \text{ then } P(y \in \mathbb{R}^k : h_z(y) = +1) > 0\}.$$

In words, S_P is the set of points x such that every linear separator whose separating hyperplane passes through x classifies a nonzero-probability region positive. The set S_P has many remarkable properties (a few of which we prove below). For instance, one can show that, when P has a density, S_P is the smallest convex set S with $P(S) = 1$, or equivalently the intersection of all convex sets S with $P(S) = 1$. These properties do not necessarily hold for distributions that do not have a density, though Witsenhausen [1968] shows they do hold for arbitrary P when $k = 1$. For our purposes, we will establish a few basic properties of S_P, stated in the following lemmas.

Lemma 7.17. For any probability measure P over $\mathbb{R}^k$, S_P is convex.

Proof. Let $x', x'' \in S_P$, $\alpha \in (0,1)$, and $x = \alpha x' + (1-\alpha)x''$. Fix any $z \in Z_k$ with $x \in \partial z$. Since x, x', and x'' are collinear, with x between x' and x'', we must have either $h_z(x') = +1$ or $h_z(x'') = +1$ (or both). Without loss of generality, suppose the former. Let $z' \in Z_k$ have $z'_i = z_i$ for all $i \le k$, and $z'_{k+1} = -\sum_{i=1}^k z'_i x'_i$; thus, the hyperplanes ∂z and $\partial z'$ are parallel, but $\partial z'$ is a translation of ∂z to satisfy $x' \in \partial z'$. For this reason, and since $h_z(x') = +1$, we have $\{y \in \mathbb{R}^k : h_{z'}(y) = +1\} \subseteq \{y \in \mathbb{R}^k : h_z(y) = +1\}$. In particular, this implies $P(y \in \mathbb{R}^k : h_z(y) = +1) \ge P(y \in \mathbb{R}^k : h_{z'}(y) = +1) > 0$, where this last inequality is due to the fact that $x' \in S_P$. Since this holds for every $z \in Z_k$ with $x \in \partial z$, we have $x \in S_P$. $\square$

Lemma 7.18. Let P be any probability measure over $\mathbb{R}^k$, and let $X \sim P$. For any measurable set $B \subseteq \mathbb{R}^k$ with $\operatorname{diam}(B) < \infty$ and $P(B) > 0$, the point $x^B = \mathbb{E}[X|X \in B]$ satisfies $x^B \in S_P$.

Proof. Fix such a set B, and let $z \in Z_k$ satisfy $x^B \in \partial z$. By linearity of expectations, this means

$$\mathbb{E}\left[z_{k+1} + \sum_{i=1}^{k} z_i X_i \middle| X \in B \right] = z_{k+1} + \sum_{i=1}^{k} z_i x_i^B = 0.$$

This immediately implies a nonzero conditional probability that $z_{k+1} + \sum_{i=1}^k z_i X_i \ge 0$, given $X \in B$. Since X has distribution P, this means

$$P\left(\left\{ x : z_{k+1} + \sum_{i=1}^{k} z_k x_i \ge 0 \right\} \middle| B \right) > 0,$$

where $P(\cdot|B) = P(\cdot \cap B)/P(B)$, as usual. Therefore,

$$\begin{aligned} P\left(x : h_z(x) = +1\right) &= P\left(x : z_{k+1} + \sum_{i=1}^{k} z_k x_i \geq 0\right) \\ &\geq P\left(\left\{x : z_{k+1} + \sum_{i=1}^{k} z_k x_i \geq 0\right\} \cap B\right) \\ &= P\left(\left\{z_{k+1} + \sum_{i=1}^{k} z_k x_i \geq 0\right\} \middle| B\right) P(B) > 0. \end{aligned}$$

□

Lemma 7.19. For any probability measure P over $\mathbb{R}^k$ that has a density (with respect to λ^k), $P(S_P) = 1$.

Proof. For the purpose of contradiction, suppose $P(\mathbb{R}^k \setminus S_P) > 0$. Since Lemma 7.17 implies S_P is convex, we know $\lambda^k(\bar{S}_P \setminus S_P) = 0$, where $\bar{S}_P$ is the closure of S_P [see e.g., Bogachev, 1998, Lemma 1.8.1]; since P has a density with respect to λ^k, the Radon-Nikodym theorem implies $P(S_P) = P(\bar{S}_P)$, and thus we have $P(\mathbb{R}^k \setminus \bar{S}_P) > 0$. Since $\mathbb{R}^k \setminus \bar{S}_P$ is an open set, and the collection of rational-radius open balls centered at rational points forms a basis for the Euclidean topology, there exists a countable collection $\{B_i\}_{i=1}^{\infty}$ of these balls such that $\bigcup_{i=1}^{\infty} B_i = \mathbb{R}^k \setminus \bar{S}_P$. In particular, each B_i has $\mathrm{diam}(B_i) < \infty$, and $B_i \cap S_P = \emptyset$. Since $P(\mathbb{R}^k \setminus \bar{S}_P) > 0$, at least one of these balls, B_i, has $P(B_i) > 0$. By Lemma 7.18, the point $x^{B_i} = \mathbb{E}[X | X \in B_i]$ satisfies $x^{B_i} \in S_P$, where $X \sim P$. But since B_i is a ball, and hence convex, we have $x^{B_i} \in B_i$, so that $B_i \cap S_P \neq \emptyset$: a contradiction. □

Lemma 7.20. For any probability measure P over $\mathbb{R}^k$ that has a density (with respect to λ^k), any $x^0 \in S_P$ has

$$\inf\{P(x : h_z(x) = +1) : z \in Z_k, x^0 \in \partial z\} > 0.$$

Proof. Fix any $x^0 \in S_P$. First note that, for any $z \in Z_k$, the vector $z' \in Z_k$ with $\forall i \leq k+1, z'_i = z_i / \|(z_1, \ldots, z_k)\|$, has $h_{z'} = h_z$ and $\partial z' = \partial z$. Thus, letting $Z'_k = \{z \in Z_k : \|(z_1, \ldots, z_k)\| = 1\}$, it suffices to show $\inf\{P(x : h_z(x) = +1) : z \in Z'_k, x^0 \in \partial z\} > 0$. Furthermore, any $z = (z_1, \ldots, z_{k+1}) \in Z'_k$ with $x^0 \in \partial z$ has $z_{k+1} = -\sum_{i=1}^{k} z_i x_i^0$. Thus, if we define a function $\tilde{z}$ mapping any $w = (w_1, \ldots, w_k) \in \mathbb{R}^k$ with $\|w\| = 1$ to the vector $\tilde{z}(w) = (w_1, \ldots, w_k, -\sum_{i=1}^{k} w_i x_i^0)$, it suffices to show $\inf\{P(x : h_{\tilde{z}(w)}(x) = +1) : w \in \mathbb{R}^k, \|w\| = 1\} > 0$.

Note that $\tilde{z}$ is continuous over $\{w \in \mathbb{R}^k : \|w\| = 1\}$. Furthermore, since P has a density with respect to λ^k, the function $z \mapsto P(x : h_z(x) = +1)$ is continuous over Z_k; thus, since compositions of continuous functions are continuous, the function $w \mapsto P(x : h_{\tilde{z}(w)}(x) = +1)$ is continuous over $\{w \in \mathbb{R}^k : \|w\| = 1\}$. The set $\{w \in \mathbb{R}^k : \|w\| = 1\}$ is a unit-radius sphere in $\mathbb{R}^k$, which is a compact set [e.g., Munkres, 2000, page 174]. Together with the extreme value theorem, the previous two facts imply that $\exists w^0 \in \mathbb{R}^k$ with $\|w^0\| = 1$ such that $P(x : h_{\tilde{z}(w^0)}(x) = +1) = \inf\{P(x : h_{\tilde{z}(w)}(x) = +1) : w \in \mathbb{R}^k, \|w\| = 1\}$. Furthermore, since $x^0 \in S_P$, and $x^0 \in \partial\tilde{z}(w^0)$ by definition of $\tilde{z}$, we must have $P(x : h_{\tilde{z}(w^0)}(x) = +1) > 0$. □

We are now ready for the proof of Theorem 7.15.

Proof of Theorem 7.15. Fix any $h \in \mathbb{C}$. We will establish that $\forall y \in \mathcal{Y}, \mathcal{P}(\{x : h(x) = y\} \cap \partial h) = 0$; note that since $\mathcal{P}(\partial h) = \sum_{y \in \mathcal{Y}} \mathcal{P}(\{x : h(x) = y\} \cap \partial h)$, this would entail $\mathcal{P}(\partial h) = 0$, which, combined with Lemma 7.12, would imply the result.

Now fix any $y \in \mathcal{Y}$, and let $p_y = \mathcal{P}(x : h(x) = y)$. In the trivial case of $p_y = 0$, certainly $\mathcal{P}(\{x : h(x) = y\} \cap \partial h) \leq p_y = 0$. For the remaining case, suppose $p_y > 0$, and define the conditional probability measure $\mathcal{P}^y(\cdot) = \mathcal{P}(\cdot|\{x : h(x) = y\}) = \mathcal{P}(\cdot \cap \{x : h(x) = y\})/p_y$; since $\mathcal{P}(\{x : h(x) = y\} \cap \partial h) \leq \mathcal{P}^y(\partial h)$, it suffices to show $\mathcal{P}^y(\partial h) = 0$. Note that, since $\mathcal{P}$ has a density p with respect to λ^k, $\mathcal{P}^y$ has a density with respect to λ^k as well: namely, $x \mapsto \mathbb{1}[h(x) = y]p(x)/p_y$. Therefore, Lemma 7.19 implies $\mathcal{P}^y(S_{\mathcal{P}^y}) = 1$, so that it suffices for us to show $S_{\mathcal{P}^y} \cap \partial h = \emptyset$.

Toward this end, fix any $x^0 \in S_{\mathcal{P}^y}$. Note that $h(x^0) = y$ (otherwise, since $\mathcal{P}^y(x : h(x) = -y) = 0$, one of the two separators parallel to that of h and passing through x^0 would witness $x^0 \notin S_{\mathcal{P}^y}$). Let $q = \inf\{\mathcal{P}^y(x : h_z(x) = +1) : z \in Z_k, x^0 \in \partial z\}$, and note that since $\mathcal{P}^y$ has a density with respect to λ^k, Lemma 7.20 implies that $q > 0$.

Let $z \in Z_k$ be such that $h_z(x^0) \neq h(x^0)$. Furthermore, let $z' \in Z_k$ have $z'_i = z_i$ for all $i \leq k$, and $z'_{k+1} = -\sum_{i=1}^k z'_i x^0_i$: that is, $\partial z'$ is parallel to ∂z, but translated so that $x^0 \in \partial z'$. Note that, since $y = h(x^0) \neq h_z(x^0)$, we have $y \sum_{i=1}^k z_i x^0_i \leq -y z_{k+1}$. Thus, since $-y z'_{k+1} = y \sum_{i=1}^k z_i x^0_i$, we have $-y z'_{k+1} \leq -y z_{k+1}$, with strict inequality if $z' \neq z$. Therefore, if $z' \neq z$, for any x with $h_z(x) = y$, since $y \sum_{i=1}^k z_i x_i \geq -y z_{k+1}$, we have $y \sum_{i=1}^k z'_i x_i = y \sum_{i=1}^k z_i x_i \geq -y z_{k+1} > -y z'_{k+1}$, so that $h_{z'}(x) = y$ as well. In particular, this implies $\{x : h(x) = y \neq h_{z'}(x)\} \subseteq \{x : h(x) = y \neq h_z(x)\}$. Thus, $\mathcal{P}(x : h(x) \neq h_z(x)) \geq \mathcal{P}^y(x : h(x) \neq h_z(x)) p_y \geq \mathcal{P}^y(x : h(x) \neq h_{z'}(x)) p_y = \mathcal{P}^y(x : h_{z'}(x) \neq y) p_y$. If $y = -1$, we have $\mathcal{P}^y(x : h_{z'}(x) \neq y) = \mathcal{P}^y(x : h_{z'}(x) = +1) \geq q$. Otherwise, $y = +1$. In this case, note that $\{x : h_{z'}(x) \neq y\} = \{x : h_{-z'}(x) = +1\} \setminus \partial z'$. Since $\lambda^k(\partial z') = 0$ and $\mathcal{P}^y$ has a density with respect to λ^k, the Radon-Nikodym theorem implies $\mathcal{P}^y(x : h_{z'}(x) \neq y) = \mathcal{P}^y(x : h_{-z'}(x) = +1)$. Furthermore, since $\partial(-z') = \partial z'$, we have $x^0 \in \partial(-z')$ as well, so that $\mathcal{P}^y(x : h_{-z'}(x) = +1) \geq q$.

Combining the above arguments, we have established that any $z \in Z_k$ with $h_z(x^0) \neq h(x^0)$ has $\mathcal{P}(x : h(x) \neq h_z(x)) \geq p_y q > 0$, so that $x^0 \notin \mathrm{DIS}(\mathrm{B}(h, p_y q/2))$; since $\partial h \subseteq \mathrm{DIS}(\mathrm{B}(h, q p_y/2))$, this implies $x^0 \notin \partial h$. This holds for all $x^0 \in S_{\mathcal{P}^y}$, so that $S_{\mathcal{P}^y} \cap \partial h = \emptyset$. $\square$

7.3.3 Proof of Theorem 7.16

Here we present the formal proof of Theorem 7.16. Again, we continue the notational conventions introduced above. Additionally, for any $r > 0$, let

$$\mathrm{B}_k(r) = \{x \in \mathbb{R}^k : \|x\| \leq r\}$$

denote the origin-centered ball of radius r. Before stating the proof, we review a basic lemma on the volume of a certain spherical segment.

Lemma 7.21. For any $r > 0$, any $\sigma > 0$ with $\sigma \leq r/(3\sqrt{k})$, and any $w = (w_1, \ldots, w_k) \in \mathbb{R}^k$ with $\|w\| = 1$,

$$\frac{\sqrt{k}}{2r}\sigma \leq \frac{\lambda^k\left(x \in \mathrm{B}_k(r) : \left|\sum_{i=1}^k w_i x_i\right| \leq \sigma\right)}{\lambda^k(\mathrm{B}_k(r))} \leq \frac{\sqrt{k}}{r}\sigma.$$

Proof. Fix r, σ, and w as in the lemma. We can express $\lambda^k\left(x \in \mathrm{B}_k(r) : \left|\sum_{i=1}^k w_i x_i\right| \geq \sigma\right) = 2C_k(r - \sigma)$, where $C_k(u)$ is the volume of a hyperspherical cap of height $u \in (0, r)$. It is known that $C_k(u) = \frac{1}{2}\lambda^k(\mathrm{B}_k(r)) I_{2\frac{u}{r} - \frac{u^2}{r^2}}\left(\frac{k+1}{2}, \frac{1}{2}\right)$ [Li, 2011], where $I_y(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^y t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a+b)}{a\Gamma(a)\Gamma(b)} \int_0^{y^a} (1 - v^{1/a})^{b-1}\mathrm{d}v$ is the regularized incomplete beta function, and Γ is the usual gamma function: $\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}\mathrm{d}t$. We therefore have $\frac{\lambda^k\left(x \in \mathrm{B}_k(r) : \left|\sum_{i=1}^k w_i x_i\right| \le \sigma\right)}{\lambda^k(\mathrm{B}_k(r))} = 1 - \frac{2C_k(r-\sigma)}{\lambda^k(\mathrm{B}_k(r))} = 1 - I_{1-\sigma^2/r^2}\left(\frac{k+1}{2}, \frac{1}{2}\right)$. Since $I_y(a,b) = 1 - I_{1-y}(b,a)$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, this equals

$$I_{\sigma^2/r^2}\left(\frac{1}{2}, \frac{k+1}{2}\right) = \frac{2\Gamma\left(\frac{k+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{k+1}{2}\right)} \int_0^{\sigma/r} \left(1 - x^2\right)^{\frac{k-1}{2}} \mathrm{d}x. \tag{7.4}$$

Since $\frac{2\Gamma\left(\frac{k+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{k+1}{2}\right)} \ge \sqrt{\frac{k}{3}}$, the right hand side of (7.4) is at least $\sqrt{\frac{k}{3}} \int_0^{\sigma/r} (1-x^2)^{\frac{k}{2}} \mathrm{d}x$. For $x \le 1/3$, $1 - x^2 \ge e^{-2x^2}$, so that when $\sigma/r \le 1/3$, the above is at least $\sqrt{\frac{\pi}{3}} \int_0^{\sigma/r} \sqrt{\frac{k}{\pi}} e^{-kx^2}\mathrm{d}x \ge \frac{1}{2}\min\left\{\frac{1}{2}, \sqrt{k}\sigma/r\right\} = \sqrt{k}\sigma/(2r)$, which proves the left inequality in the lemma statement.

Furthermore, since $\frac{2\Gamma\left(\frac{k+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{k+1}{2}\right)} \le \sqrt{k}$, the right hand side of (7.4) is at most $\sqrt{k}\int_0^{\sigma/r} (1-x^2)^{\frac{k-1}{2}} \mathrm{d}x$; since $\sigma/r \le 1$, any $x \in (0, \sigma/r)$ has $(1-x^2)^{\frac{k-1}{2}} \le 1$, so that $\sqrt{k}\int_0^{\sigma/r} (1-x^2)^{\frac{k-1}{2}} \mathrm{d}x \le \sqrt{k}\sigma/r$, which proves the remaining inequality in the lemma statement. □

With this lemma in hand, we next prove Theorem 7.16.

Proof of Theorem 7.16. We use an argument based on the analyses of linear separators under uniform distributions by Hanneke [2007b] and Balcan, Hanneke, and Vaughan [2010]. For simplicity, we make relatively little effort to optimize the constant factors in this analysis.

Let $Z'_k = \{z \in Z_k : \sum_{i=1}^k z_i^2 = 1\}$, and note that, as explained in the proof of Lemma 7.20, we can equivalently express $\mathbb{C} = \{h_z : z \in Z'_k\}$. Now fix any $z \in Z'_k$ and $\mathcal{P}$ as in the theorem statement. By Corollary 7.10, it suffices to show that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon))) = O(\varepsilon)$, so we focus on proving this. Let $x^0 \in \mathrm{supp}(p) \cap \partial z$ be a continuity point of p. To simplify the argument, let us suppose x^0 is the origin point

$(0,\dots,0)$; we lose no generality by this assumption, since we can obtain the general case by applying a translation to $\mathbb{R}^k$ (inducing the desired $\mathcal{P}$ distribution), while altering the $(k+1)^{\mathrm{th}}$ entry of z to maintain that ∂z passes through the translated x^0. This transformation does not alter the hypothesis class $\mathbb{C}$, and preserves the value of $\theta_{h_z}(\varepsilon)$. In particular, note that taking x^0 as the origin implies $z_{k+1} = 0$.

Define $p_0 = p(x^0)/2$, and note $p_0 > 0$ since $x^0 \in \mathrm{supp}(p)$. Let $r > 0$ be chosen so that $\inf_{x \in \mathrm{B}_k(r)} p(x) \geq p_0$; such an r exists due to continuity of p at x^0. Let $p_\pm = \min_{y \in \mathcal{Y}} \mathcal{P}(x \in \mathrm{B}_k(r) : h_z(x) = y)$, and note that $p_\pm \geq p_0 \min_{y \in \mathcal{Y}} \lambda^k(x \in \mathrm{B}_k(r) : h_z(x) = y) = p_0 \lambda^k(\mathrm{B}_k(r))/2$, where this last equality is due to $x^0 \in \partial z$ (i.e., h_z bisects $\mathrm{B}_k(r)$); in particular, $p_\pm > 0$. Now fix any $\varepsilon \in (0, p_\pm)$, and consider any $z' \in Z'_k$ with $h_{z'} \in \mathrm{B}(h_z, \varepsilon)$.

Our first task is to lower bound $\mathcal{P}(x : h_z(x) \neq h_{z'}(x))$ by a function of z and z', so that we may characterize properties satisfied by all elements of $\mathrm{B}(h_z, \varepsilon)$. Toward this end, we will focus on the neighborhood $\mathrm{B}_k(r)$ of x^0, since (as we show) any differences between $h_{z'}$ and h_z will be reflected to some extent there. Let $\mathcal{P}_0$ be a probability measure on $\mathcal{X}$ such that, for any measurable set $A \subseteq \mathcal{X}$, $\mathcal{P}_0(A) = \lambda^k(A \cap \mathrm{B}_k(r))/\lambda^k(\mathrm{B}_k(r))$; that is, $\mathcal{P}_0$ is a uniform distribution in $\mathrm{B}_k(r)$. Since $p(x) \geq p_0$ for all $x \in \mathrm{B}_k(r)$, and p is a density with respect to λ^k, we must have

$$\begin{aligned}\mathcal{P}(x : h_z(x) \neq h_{z'}(x)) &\geq p_0 \lambda^k\left(x \in \mathrm{B}_k(r) : h_z(x) \neq h_{z'}(x)\right)\\ &= p_0 \lambda^k(\mathrm{B}_k(r)) \mathcal{P}_0(x : h_z(x) \neq h_{z'}(x)).\end{aligned}$$

We are therefore interested in lower bounds on $\mathcal{P}_0(x : h_z(x) \neq h_{z'}(x))$.

Since $\varepsilon < p_\pm$, we know $\min_{y \in \mathcal{Y}} \mathcal{P}(x \in \mathrm{B}_k(r) : h_{z'}(x) = h_z(x) = y) > 0$. Since $\mathcal{P}$ has a density with respect to λ^k, the Radon-Nikodym theorem implies $\min_{y \in \mathcal{Y}} \lambda^k(x \in \mathrm{B}_k(r) : h_{z'}(x) = h_z(x) = y) > 0$; in particular, letting $q_{z'} = \min_{y \in \mathcal{Y}} \mathcal{P}_0(x : h_{z'}(x) = y)$, we have $q_{z'} > 0$. Furthermore, again because $x^0 \in \partial z$, we have $\min_{y \in \mathcal{Y}} \mathcal{P}_0(x : h_z(x) = y) = 1/2$. Based on this, we can already identify one basic lower bound, simply by noting that

$$\mathcal{P}_0(x : h_z(x) \neq h_{z'}(x)) \geq \max_{y \in \mathcal{Y}} \mathcal{P}_0(x : h_z(x) = y \neq h_{z'}(x))$$
$$\geq \max_{y \in \mathcal{Y}} \mathcal{P}_0(x : h_z(x) = y) - \mathcal{P}_0(x : h_{z'}(x) = y) = \frac{1}{2} - q_{z'}. \qquad (7.5)$$

We can also obtain a less-obvious lower bound based on the *angle* between the separating hyperplanes of h_z and $h_{z'}$, as follows. For any $z^a, z^b \in Z'_k$, let $\alpha(z^a, z^b) = \arccos(\sum_{i=1}^{k} z_i^a z_i^b) \in [0, \pi]$ denote the angle between the vectors $(z_1^a, \ldots, z_k^a)$ and $(z_1^b, \ldots, z_k^b)$. Now consider a new vector $\bar{z} = (\bar{z}_1, \ldots, \bar{z}_{k+1})$ such that $\bar{z}_i = z'_i$ for $i \leq k$, and $\bar{z}_{k+1} = 0$; the hyperplane $\partial\bar{z}$ is parallel to $\partial z'$, and thus $\alpha(z, \bar{z}) = \alpha(z, z')$, but the intercept of $\bar{z}$ differs from that of z' so that the separating hyperplane passes through x^0: that is, $x^0 \in \partial\bar{z}$. Since $\{x : h_z(x) \neq h_{\bar{z}}(x)\} \subseteq \{x : h_z(x) \neq h_{z'}(x)\} \cup \{x : h_{\bar{z}}(x) \neq h_{z'}(x)\}$, a union bound implies

$$\mathcal{P}_0(x : h_z(x) \neq h_{z'}(x))$$
$$\geq \mathcal{P}_0(x : h_z(x) \neq h_{\bar{z}}(x)) - \mathcal{P}_0(x : h_{\bar{z}}(x) \neq h_{z'}(x)). \quad (7.6)$$

Note that, since the hyperplanes $\partial z'$ and $\partial\bar{z}$ are parallel, and $x^0 \in \partial\bar{z}$, we have $\mathcal{P}_0(x : h_{\bar{z}}(x) \neq h_{z'}(x)) = |\mathcal{P}_0(x : h_{\bar{z}}(x) = +1) - \mathcal{P}_0(x : h_{z'}(x) = +1)| = \frac{1}{2} - q_{z'}$. Furthermore, since $x^0 \in \partial\bar{z}$ and $x^0 \in \partial z$, by considering the projection of $\mathcal{P}_0$ onto the subspace spanned by $(z_1, \ldots, z_k)$ and $(\bar{z}_1, \ldots, \bar{z}_k)$, we see that $\mathcal{P}_0(x : h_z(x) \neq h_{\bar{z}}(x)) = \alpha(z, \bar{z})/\pi = \alpha(z, z')/\pi$. Combining these observations with (7.6) and (7.5), we have

$$\mathcal{P}_0(x : h_z(x) \neq h_{z'}(x)) \geq \max\left\{\frac{\alpha(z, z')}{\pi} - \left(\frac{1}{2} - q_{z'}\right), \frac{1}{2} - q_{z'}\right\}$$
$$\geq \frac{1}{2}\left(\frac{\alpha(z, z')}{\pi} - \left(\frac{1}{2} - q_{z'}\right)\right) + \frac{1}{2}\left(\frac{1}{2} - q_{z'}\right) = \frac{\alpha(z, z')}{2\pi}.$$

Combining this with (7.5), we generally have

$$\mathcal{P}_0(x : h_z(x) \neq h_{z'}(x)) \geq \max\left\{\frac{\alpha(z, z')}{2\pi}, \frac{1}{2} - q_{z'}\right\}. \qquad (7.7)$$

Letting

$$Z_\varepsilon = \left\{z' \in Z'_k : \max\left\{\frac{\alpha(z, z')}{2\pi}, \frac{1}{2} - q_{z'}\right\} \leq \varepsilon/(p_0 \lambda^k(\mathrm{B}_k(r)))\right\},$$

we have thus proven that $\mathrm{B}(h_z, \varepsilon) \subseteq \{h_{z'} : z' \in Z_\varepsilon\}$. It remains only to characterize the region of disagreement of this latter set. To focus on nontrivial cases, suppose $\varepsilon < p_0 \lambda^k(\mathrm{B}_k(r))/12$. Let z^+ and z^- be the two elements of Z_ε with $z_i^+ = z_i^- = z_i$ for all $i \leq k$, and with z_{k+1}^+ and z_{k+1}^- set so that $\mathcal{P}_0(x : h_{z^+}(x) = +1) = q_{z^+} = \mathcal{P}_0(x : h_{z^-}(x) = -1) = q_{z^-} = \frac{1}{2} - \varepsilon/(p_0 \lambda^k(\mathrm{B}_k(r)))$; in particular, note that $z_{k+1}^+ = -z_{k+1}^- \in (-r, r)$. Since the value of $q_{z'}$ is completely determined by $|z'_{k+1}|$, and is strictly decreasing in this quantity for $|z'_{k+1}| \in (0, r)$, every $z' \in Z_\varepsilon$ satisfies $|z'_{k+1}| \leq |z_{k+1}^+|$. Therefore, for any $z' \in Z_\varepsilon$, and for $\tilde{z} = (z_1, \ldots, z_k, z'_{k+1})$, we have

$$\begin{aligned}\mathrm{DIS}(\{h_{z'}, h_z\}) &\subseteq \mathrm{DIS}(\{h_z, h_{\tilde{z}}\}) \cup \mathrm{DIS}(\{h_{\tilde{z}}, h_{z'}\}) \\ &\subseteq \mathrm{DIS}(\{h_{z^+}, h_{z^-}\}) \cup \mathrm{DIS}(\{h_{\tilde{z}}, h_{z'}\}).\end{aligned}$$

Furthermore, letting $z'^+ = (z'_1, \ldots, z'_k, z_{k+1}^+)$ and $z'^- = (z'_1, \ldots, z'_k, z_{k+1}^-)$, we have that $\mathrm{DIS}(\{h_{\tilde{z}}, h_{z'}\}) \subseteq \mathrm{DIS}(\{h_{z^+}, h_{z^-}\}) \cup \mathrm{DIS}(\{h_{z^+}, h_{z'^+}\}) \cup \mathrm{DIS}(\{h_{z^-}, h_{z'^-}\})$ (this becomes clear when one considers the projection onto the space spanned by the vectors $(z'_1, \ldots, z'_k)$ and $(z_1, \ldots, z_k)$). Thus, we have that

$$\mathrm{DIS}(\{h_{z'}, h_z\}) \subseteq \mathrm{DIS}(\{h_{z^+}, h_{z^-}\}) \cup \mathrm{DIS}(\{h_{z^+}, h_{z'^+}\}) \cup \mathrm{DIS}(\{h_{z^-}, h_{z'^-}\}).$$

Applying this to every $z' \in Z_\varepsilon$, we find that, letting $Z_\varepsilon^+ = \left\{h_{z'} : z' \in Z'_k, z'_{k+1} = z_{k+1}^+, \frac{\alpha(z', z)}{2\pi} \leq \varepsilon/(p_0 \lambda^k(\mathrm{B}_k(r)))\right\}$ and $Z_\varepsilon^- = \left\{h_{z'} : z' \in Z'_k, z'_{k+1} = z_{k+1}^-, \frac{\alpha(z', z)}{2\pi} \leq \varepsilon/(p_0 \lambda^k(\mathrm{B}_k(r)))\right\}$,

$$\begin{aligned}&\mathrm{DIS}\left(\{h_{z'} : z' \in Z_\varepsilon\}\right) = \bigcup_{z' \in Z_\varepsilon} \mathrm{DIS}(\{h_{z'}, h_z\}) \\ &\subseteq \mathrm{DIS}(\{h_{z^+}, h_{z^-}\}) \cup \mathrm{DIS}\left(\left\{h_{z'} : z' \in Z_\varepsilon^+\right\}\right) \cup \mathrm{DIS}\left(\{h_{z'} : z' \in Z_\varepsilon^-\}\right).\end{aligned} \tag{7.8}$$

Note that the region $\mathrm{DIS}(\{h_{z^+}, h_{z^-}\})$ is simply a fixed-width slab around ∂z: namely, $\left\{x : \left|\sum_{i=1}^k z_i x_i\right| \leq |z_{k+1}^+|\right\}$. Furthermore, we can bound the size of $|z_{k+1}^+|$, as follows. By monotonicity of measures and Lemma 7.21, we have $\mathcal{P}_0\left(\mathrm{DIS}(h_{z^+}, h_{z^-})\right) = \mathcal{P}_0\left(x : \left|\sum_{i=1}^k z_i x_i\right| \leq |z_{k+1}^+|\right) \geq \mathcal{P}_0\left(x : \left|\sum_{i=1}^k z_i x_i\right| \leq \min\left\{|z_{k+1}^+|, r/(3\sqrt{k})\right\}\right) \geq \min\left\{\frac{\sqrt{k}}{2r}|z_{k+1}^+|, \frac{1}{6}\right\}$.

Since $\mathcal{P}_0(\mathrm{DIS}(\{h_{z^+}, h_{z^-}\})) = \mathcal{P}_0(\mathrm{DIS}(\{h_{z^+}, h_z\})) + \mathcal{P}_0(\mathrm{DIS}(\{h_{z^-}, h_z\})) = \left(\frac{1}{2} - q_{z^-}\right) + \left(\frac{1}{2} - q_{z^+}\right) = 2\varepsilon/(p_0\lambda^k(\mathrm{B}_k(r))) < 1/6$, we have that $\frac{\sqrt{k}}{2r}|z^+_{k+1}| \leq 2\varepsilon/(p_0\lambda^k(\mathrm{B}_k(r)))$; letting $t_1 = \frac{4r}{\sqrt{k}p_0\lambda^k(\mathrm{B}_k(r))}$, this implies

$$|z^+_{k+1}| \leq t_1\varepsilon. \tag{7.9}$$

Next, we bound the remaining region in (7.8). Consider any $z', z'' \in Z'_k$ with $z'_{k+1} = z''_{k+1} \in (-r/\sqrt{2}, r/\sqrt{2})$ and $\alpha(z', z'') \leq \pi/4$. Some basic trigonometry in the space spanned by $(z'_1, \ldots, z'_k)$ and $(z''_1, \ldots, z''_k)$ reveals that there exists a point $x \in \partial z' \cap \partial z''$ with $\|x\| \leq |z'_{k+1}|/\cos(\alpha(z', z'')/2) \leq \sqrt{2}|z'_{k+1}| < r$, so that $x \in \mathrm{B}_k(r)$, and therefore $x \in \mathrm{supp}(p)$ as well; based on this, a little more trigonometry reveals that

$$\sup\left\{\left|z''_{k+1} + \sum_{i=1}^{k} z''_i x'_i\right| : x' \in \mathrm{supp}(p), h_{z'}(x') \neq h_{z''}(x')\right\}$$
$$\leq \mathrm{diam}(\mathrm{supp}(p))\frac{\sin(\alpha(z', z''))}{\cos(\alpha(z', z''))} \leq \mathrm{diam}(\mathrm{supp}(p))\sqrt{2}\alpha(z', z''),$$

where this last inequality follows from the facts that $\cos(\alpha(z', z'')) \geq 1/\sqrt{2}$ and $\sin(\alpha(z', z'')) \leq \alpha(z', z'')$.

Now let us apply these observations to the specific vectors in Z^+_ε and Z^-_ε. Since $\varepsilon \leq p_0\lambda^k(\mathrm{B}_k(r))/8$, we are guaranteed every $z' \in Z^+_\varepsilon$ has $\alpha(z', z^+) \leq \pi/4$, and likewise every $z' \in Z^-_\varepsilon$ has $\alpha(z', z^-) \leq \pi/4$. Furthermore, since $\varepsilon < p_0\lambda^k(\mathrm{B}_k(r))/12 < r/(\sqrt{2}t_1)$, (7.9) implies $z^-_{k+1}, z^+_{k+1} \in (-r/\sqrt{2}, r/\sqrt{2})$. Thus, letting $t_2 = \frac{\mathrm{diam}(\mathrm{supp}(p))\sqrt{8}\pi}{p_0\lambda^k(\mathrm{B}_k(r))}$, combining the above argument with the bound on $\alpha(z', z)$ from the definitions of Z^+_ε and Z^-_ε, and noting $\alpha(z', z^+) = \alpha(z', z^-) = \alpha(z', z)$, we have

$$\mathrm{DIS}(\{h_{z'} : z' \in Z^+_\varepsilon\}) \cap \mathrm{supp}(p) = \bigcup_{z' \in Z^+_\varepsilon} \mathrm{DIS}(\{h_{z'}, h_{z^+}\}) \cap \mathrm{supp}(p)$$
$$\subseteq \left\{x \in \mathrm{supp}(p) : \left|z^+_{k+1} + \sum_{i=1}^{k} z^+_i x_i\right| \leq t_2\varepsilon\right\},$$

and similarly

$$\begin{aligned}&\mathrm{DIS}(\{h_{z'} : z' \in Z_\varepsilon^-\}) \cap \mathrm{supp}(p) \\ &\qquad\subseteq \left\{x \in \mathrm{supp}(p) : \left| z_{k+1}^- + \sum_{i=1}^k z_i^- x_i \right| \leq t_2 \varepsilon \right\}.\end{aligned}$$

Combining this with the fact that ∂z^+ and ∂z^- are parallel to ∂z, and plugging into (7.8), together with (7.9), we have

$$\begin{aligned}&\mathrm{DIS}(\{h_{z'} : z' \in Z_\varepsilon\}) \cap \mathrm{supp}(p) \\ &\qquad\subseteq \left\{x \in \mathrm{supp}(p) : \left| \sum_{i=1}^k z_i x_i \right| \leq |z_{k+1}^+| + t_2 \varepsilon \right\} \\ &\qquad\subseteq \left\{x \in \mathrm{supp}(p) : \left| \sum_{i=1}^k z_i x_i \right| \leq (t_1 + t_2)\, \varepsilon \right\}.\end{aligned}$$

Let $p_{\max} = \sup_{x \in \mathrm{supp}(p)} p(x)$, and note that since p is bounded, $p_{\max} < \infty$. Let $r' = \sup_{x \in \mathrm{supp}(p)} \|x\|$, and note that $0 < r \leq r' \leq \mathrm{diam}(\mathrm{supp}(p)) < \infty$. From the above arguments, we conclude that

$$\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon)) \cap \mathrm{supp}(p) \subseteq \left\{x \in \mathrm{B}_k(r') : \left| \sum_{i=1}^k z_i x_i \right| \leq (t_1 + t_2)\varepsilon \right\},$$

so that

$$\begin{aligned}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon))) &\leq \mathcal{P}\left(x \in \mathrm{B}_k(r') : \left| \sum_{i=1}^k z_i x_i \right| \leq (t_1 + t_2)\varepsilon \right) \\ &\leq p_{\max} \lambda^k \left(x \in \mathrm{B}_k(r') : \left| \sum_{i=1}^k z_i x_i \right| \leq (t_1 + t_2)\varepsilon \right).\end{aligned}$$

For any $\varepsilon < r'/(3\sqrt{k}(t_1 + t_2))$, Lemma 7.21 implies this is at most $p_{\max} \lambda^k(\mathrm{B}_k(r')) \frac{\sqrt{k}}{r'}(t_1 + t_2)\varepsilon$. Since this is $O(\varepsilon)$, Corollary 7.10 implies $\theta_{h_z}(\varepsilon) = O(1)$. □

7.3.4 General Analysis: Smooth Functions

This subsection describes general sufficient conditions on $\mathbb{C}$ and $\mathcal{P}$ for $\theta_h(\varepsilon) = O(1)$ to hold for all $h \in \mathbb{C}$. The results in this subsection originate in the work of Friedman [2009], who presents a very general analysis of classes $\mathbb{C}$ specified by thresholding a smooth function f^z that is itself smoothly parametrized by a finite-dimensional parameter vector z. One natural motivation for considering smooth functions is that any sufficiently smooth function will be approximately linear in any small enough neighborhood. Since f^z is also smooth in the parameter vector z, we can essentially think of the hypothesis class as being, at least in any small-enough regions (in both $\mathcal{X}$ and $\mathbb{C}$), well-represented by the class of linear separators. As we have seen above, linear separators h have bounded disagreement coefficients under a few simple conditions on h and $\mathcal{P}$, so we might expect the same to be true of these smooth functions, since they are locally well-approximated by linear separators. There are some additional technical conditions and arguments needed to make this intuitive motivation formally correct. For instance, we need conditions guaranteeing that small neighborhoods in $\mathbb{C}$ correspond to small neighborhoods in the parameter space and vice versa, in order for the above argument regarding smoothness of f^z in z to be valid. The interested reader is referred to the original work of Friedman [2009] for the details of how these conditions come into the analysis. The formal set of conditions is stated as follows.

Condition 7.22. Suppose, for some $k, m \in \mathbb{N}$,

- $\mathcal{X}$ is a compact full-dimensional subset of $\mathbb{R}^k$.
- Z is an open subset of $\mathbb{R}^m$.
- $\mathcal{P}$ has a continuous (on $\mathcal{X}$) strictly-positive density function p (with respect to λ^k on $\mathcal{X}$).
- $f : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ is a function with continuous gradient (also denote $f_x(z) = f^z(x) = f(x, z)$).
- $\mathbb{C} = \{h_z : z \in Z\}$, where we define $h_z(x) = \mathrm{sign}(f(x, z))$ for all $x \in \mathcal{X}, z \in Z$.
- $\forall x \in \mathcal{X}, \forall z \in Z,\ f(x, z) = 0 \Rightarrow \|\nabla f^z(x)\| > 0$ (called the *transversality* condition).

- $\forall z \in Z, \forall v \in \mathbb{R}^m \setminus \{0^m\}, \exists x \in \mathcal{X}$ with $f(x,z) = 0$ and $|v \cdot \nabla f_x(z)| > 0$ (called the *non-degeneracy* condition).
- For any $z \in Z$ and $z' \in \text{Closure}(Z)$, $h_z = h_{z'} \Rightarrow z = z'$ (called the *clone-free* condition).

Friedman [2009] proves the following theorem for $\mathbb{C}$ and $\mathcal{P}$ satisfying Condition 7.22.

Theorem 7.23. Under Condition 7.22, $\forall h \in \mathbb{C}$, $\theta_h(\varepsilon) = O(1)$.

Examples of hypothesis classes satisfying these conditions on $\mathbb{C}$ include balls and axis-aligned ellipsoids (excluding those of zero-measure interior). However, Condition 7.22 is too restrictive to allow many common hypothesis classes, including linear separators and axis-aligned rectangles. To address this, Friedman [2009] additionally presents a more general result, which relaxes the requirement that Z be an open set, instead allowing Z to be any sufficiently-smooth manifold with no boundary (where the gradient in the non-degeneracy condition is defined relative to the manifold, and v is from the tangent space to the manifold at z). This more general condition *is* satisfied by the class of linear separators (where the manifold Z is precisely the set Z_k' from the proof of Theorem 7.16). He also extends the result to allow functions $f(x,z)$ that have some limited nondifferentiable points, and thereby includes such classes as rectangles and other polytopes with a bounded number of faces (excluding those with zero-measure interior). Friedman [2009] also studies important generalizations of the conditions on $\mathcal{X}$ (equivalently, on $\mathcal{P}$). In particular, rather than requiring $\mathcal{X}$ to be a full-dimensional subset of $\mathbb{R}^k$, it suffices for $\mathcal{X}$ to be a compact subset of a sufficiently-smooth manifold, where $\mathcal{X}$ has dimension equal to that of the manifold. This generalization extends Theorem 7.23 to many other distributions: for instance, the uniform distribution on a unit sphere. For simplicity, I have only included the formal details of the basic (full-dimensional) conditions here, and refer the reader to the original work of Friedman [2009] for the details of these generalizations.

The analysis of Friedman [2009] in fact provides a more detailed result than Theorem 7.23. It further establishes that, under these conditions, $\limsup_{\varepsilon \to 0} \mathcal{P}(\text{DIS}(\text{B}(h_z, \varepsilon)))/\varepsilon \lesssim m^{3/2}$. This was recently refined by Mahalanabis [2011, 2012] to $\limsup_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon)))/\varepsilon \lesssim m$. This latter result is in fact tight (up to constants) for some scenarios satisfying Condition 7.22 (though not for others; see the results for linear separators below). A particularly simple example is the set $\mathbb{C}$ of unions of i disjoint closed intervals of nonzero width, under the uniform distribution on $[0, 1]$. Formally, a classifier h in this class is specified by values $\{z_i\}_{i=1}^{2i}$, where $0 < z_1 < \cdots < z_{2i} < 1$, and for $x \in [0, 1]$, $h(x) = +1$ if $x \in \bigcup_{j=1}^{i} [z_{2j-1}, z_{2j}]$, and otherwise $h(x) = -1$. In this case, f can be specified by a polynomial with roots at the $2i$ interval boundaries, so that $m = 2i$. For any $h \in \mathbb{C}$ and any sufficiently small ε (namely, ε less than half the width of the smallest contiguous region in $[0, 1]$ on which h is constant), $\mathrm{DIS}(\mathrm{B}(h, \varepsilon))$ is the union of $2i$ disjoint semi-closed intervals of width 2ε centered at the $2i$ decision boundary points, so that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon))) = 4i\varepsilon = 2m\varepsilon$. Thus, $\limsup\limits_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))/\varepsilon = 2m$. Friedman [2009] and Mahalanabis [2011, 2012] provide other examples as well. Note that, if we are only interested in bounding the $(\mathbb{C}, \mathcal{P}, f^\star)$-dependent constant factors in the *asymptotic* behavior of the label complexity of CAL and RobustCAL as $\varepsilon \to 0$, it is clear from the proofs of Theorems 5.1, 5.4, 6.8, and 6.9 that the quantity $\limsup_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, \varepsilon)))/\varepsilon$ can be used in place of $\theta(\cdot)$ in those results (if it is finite). In this sense, these bounds on $\limsup_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_z, \varepsilon)))/\varepsilon$ have interesting direct implications for the asymptotic analysis of the label complexity of active learning.

7.4 Detailed Analyses under Specific Conditions

There are many specific classes $\mathbb{C}$ and distributions $\mathcal{P}$ for which the disagreement coefficient has been studied in detail. We review a few such results here. For context, we also survey some of the history of the analyses of each of these classes. For brevity, the proofs of these results are omitted; the interested reader is referred to the respective cited original sources for a proof of each result.

Linear Separators: Perhaps the most well-studied hypothesis class in the active learning literature is the class of k-dimensional linear separators (Example 3). As most of the heuristic active learning methods in the empirically-driven machine learning literature are built around this hypothesis class, a thorough understanding of the capabilities of active learning methods based on it is clearly desirable.

The formal analysis of active learning with linear separators has roots in the early work on learning with membership queries [e.g., Eisenberg, 1992], but the study of the label complexity of learning linear separators in the active learning model studied here was essentially initiated by Freund, Seung, Shamir, and Tishby [1997]. They studied a slightly different setting in which the target function is also considered a random variable with a known distribution; within that setting, they were able to show that, when $\mathcal{P}$ is the uniform distribution in the unit ball, and the distribution of $f^\star$ is a certain uniform distribution over nearly-balanced linear separators, in the realizable case, a slight modification of an algorithm known as *Query By Committee* achieves a label complexity $\lesssim k \log(k/\varepsilon\delta)$. In the specific case of homogeneous linear separators, due to the symmetries of the uniform distributions, this label complexity also holds for the present model, in which the target function is considered nonrandom (in this case, randomly rotating the instance space creates the same effect as having a random $f^\star$). Dasgupta, Kalai, and Monteleoni [2005, 2009] later proved this label complexity bound for homogeneous linear separators under a uniform distribution also holds for a certain Perceptron-based active learning algorithm (with slightly worse logarithmic factors). The algorithm essentially queries the labels of points relatively close to a current hypothesized decision boundary, and updates that hypothesis using a modified Perceptron rule.

The work of Balcan, Beygelzimer, and Langford [2006], which introduced the general A^2 active learning strategy (see Section 5.3), also studied the problem of learning homogeneous linear separators under the uniform distribution, and found that when $\nu \lesssim \varepsilon/\sqrt{k}$, the algorithm achieves a label complexity $\lesssim k^2 \text{polylog}(k/\varepsilon\delta)$. Implicit in this analysis is an argument that, in the realizable case, CAL achieves a label complexity $\lesssim k^{3/2}\text{polylog}(k/\varepsilon\delta)$ for this problem. Later, using a method specifically tailored to learning linear separators, Balcan, Broder, and Zhang [2007] extended the ideas of Dasgupta, Kalai, and Monteleoni [2005, 2009] to noisy settings; their technique is similar to the method of Dasgupta, Kalai, and Monteleoni [2005, 2009], except deferring hypothesis updates until obtaining a larger number of labeled points near the current hypothesized separator. Under Condition 2.3, for this problem they show label complexity bounds nearly matching the lower bound (4.2) of Theorem 4.3 up to logarithmic factors. This work also included an explicit analysis of the label complexity of (a relaxation of) CAL for the noise-free version of this problem, finding an upper bound on the label complexity that is $\lesssim k^{3/2}\mathrm{polylog}(k/\varepsilon\delta)$, thus making explicit the earlier implicit argument of Balcan, Beygelzimer, and Langford [2006].

The analysis of $\theta_h(\varepsilon)$ for the class of linear separators was initiated in the original work of Hanneke [2007b], where it was shown that for $\mathbb{C}$ the class of homogeneous linear separators, and $\mathcal{P}$ uniform on the surface of the unit sphere, any $h \in \mathbb{C}$ has $(1/4)\min\{\pi\sqrt{k}, 1/\varepsilon\} \leq \theta_h(\varepsilon) \leq \min\{\pi\sqrt{k}, 1/\varepsilon\}$ (in fact, an upper bound of $8\pi\sqrt{k}$ can be extracted from the proof of Theorem 7.16 above). Composing this result with Theorem 5.1 and Theorem 5.4 yields label complexity bounds for CAL and RobustCAL, respectively. The bounds so-obtained are sometimes slightly worse than those obtained for the best among the methods mentioned above, which are expressly designed for this scenario. Specifically, for CAL in the realizable case, Theorem 5.1 provides an upper bound on the label complexity having dependence on k larger by a factor $\sqrt{k}$ compared to the bounds of Freund, Seung, Shamir, and Tishby [1997] and Dasgupta, Kalai, and Monteleoni [2005, 2009]; this agrees with the findings of Balcan, Beygelzimer, and Langford [2006] and Balcan, Broder, and Zhang [2007] on the label complexity of CAL. It is presently not known whether this extra factor of $\sqrt{k}$ is truly present in the label complexity of CAL, or whether it is merely a limitation of the analysis. Under Condition 2.3, Theorem 5.4 provides an upper bound on the label complexity of RobustCAL, which also has dependence on k larger by a factor $\sqrt{k}$ compared to the aforementioned bound for the method of Balcan, Broder, and Zhang [2007]. However, Theorem 5.4 also provides a general upper bound $\lesssim k^{3/2}\left(\frac{\nu^2}{\varepsilon^2}+1\right)\mathrm{polylog}(1/\varepsilon\delta)$ for

RobustCAL in this scenario, which improves over the aforementioned result of Balcan, Beygelzimer, and Langford [2006] by a factor $\sqrt{k}$. This type of bound (with $k^{3/2}$ dependence on k) was first established by Dasgupta, Hsu, and Monteleoni [2007] (for a different algorithm), and remains the best known bound for this problem (for *any* method) expressed purely in terms of ν, ε, δ, and k.

The above bound on $\theta_h(\varepsilon)$ for homogeneous linear separators was later generalized by Balcan, Hanneke, and Vaughan [2010] to include non-homogeneous linear separators. Specifically, one can extract from the details of their proof (of a related result) that, for $\mathbb{C}$ the class of all linear separators and $\mathcal{P}$ the uniform distribution on the unit sphere, every $h \in \mathbb{C}$ has $\theta_h(0) \leq 4\pi\sqrt{k}/\min_{y\in\mathcal{Y}} \mathcal{P}(x : h(x) = y)$.

All of the above analyses hold for $\mathcal{P}$ the uniform distribution on the unit sphere. There are additionally several results known for other distributions. Building on earlier work of El-Yaniv and Wiener [2010] in the related selective classification setting, El-Yaniv and Wiener [2012] proved that when $\mathcal{P}$ is any mixture of a finite constant number of multivariate normal distributions with diagonal covariance matrices of full rank, the label complexity of CAL in the realizable case has asymptotic dependence on ε at most $O\left((\log(1/\varepsilon))^{(k^2+3)/2}\log\log(1/\varepsilon)\right)$. Furthermore, combined with the lower bound of Theorem 5.2 on the number of labels requested by CAL in terms of the disagreement coefficient, this upper bound is also an upper bound on the disagreement coefficient: that is, $\theta_h(\varepsilon) = O\left((\log(1/\varepsilon))^{(k^2+3)/2}\log\log(1/\varepsilon)\right)$ (see Section 8.4 for a slightly sharper bound based on a more direct application of the work of El-Yaniv and Wiener, 2010). El-Yaniv and Wiener [2012] further showed that this upper bound on the label complexity of CAL is nearly tight in terms of its asymptotic dependence on ε, by proving a lower bound (holding for $\mathcal{P}$ a multivariate standard normal) of $\Omega\left((\log(1/\varepsilon))^{(k-1)/2}\right)$ on the number of queries made by CAL among $\Omega(1/\varepsilon)$ samples, though clearly ε must be very small for this form of the lower bound to be informative (e.g., certainly $\varepsilon < 2^{-(k-1)/2}$ is required). The technique of El-Yaniv and Wiener [2010] that enabled El-Yaniv and Wiener [2012] to obtain this upper bound on the label complexity of CAL is quite general, and complements the disagreement coefficient analysis in interesting ways; we discuss it in more detail in Chapter 8.

In a somewhat different direction, Dekel, Gentile, and Sridharan [2010] study the performance of a certain stream-based active learning algorithm under arbitrary $\mathcal{P}$ (for $\mathcal{X}$ a bounded subset of $\mathbb{R}^k$) but under a very special case of Condition 2.3. Specifically, they suppose $f^\star = h_z \in \mathbb{C}$, and the function η satisfies $2\eta(x) - 1 = \sum_{i=1}^k z_i x_i$. Under these conditions, they find a label complexity matching the asymptotic dependence on ε in the lower bound (4.2) of Theorem 4.3 up to logarithmic factors, though their dependences on k and δ are slightly worse. Their algorithm also has the advantage of being computationally efficient. It is not presently known whether the lower bound of (4.2) holds under these conditions, nor has the performance of RobustCAL been lower bounded under these specific conditions.

Axis-aligned Rectangles: Another class $\mathbb{C}$ that has been studied to some extent in the literature is *axis-aligned rectangles*. Specifically, for $\mathcal{X} = \mathbb{R}^k$ for some $k \in \mathbb{N}$, the class $\mathbb{C}$ of axis-aligned rectangles is the set of classifiers $\{h_z : z \in \mathbb{R}^{2k}\}$, where for $z = (z_1, \ldots, z_{2k}) \in \mathbb{R}^{2k}$ and $x = (x_1, \ldots, x_k) \in \mathcal{X}$, $h_z(x) = \mathbb{1}^{\pm}_{\times_{i=1}^k [z_{2i-1}, z_{2i}]}(x) = 2\prod_{i=1}^k \mathbb{1}_{[z_{2i-1}, z_{2i}]}(x_i) - 1$. The VC dimension d of this class is $2k$. When $\mathcal{P}$ is a product distribution with a density, Hanneke [2007a] found that a certain noise-robust halving-style active learning algorithm achieves a label complexity that, if $p = \mathcal{P}(x : f^\star(x) = +1) > 5\nu$, is $\lesssim \frac{k^3}{p}\left(\frac{\nu^2}{\varepsilon^2} + 1\right) \operatorname{polylog}\left(\frac{k}{\varepsilon\delta p}\right)$. The approach to proving this result employs a technique that is quite general, and we discuss it in more detail in Chapter 8.

As for the value of the disagreement coefficient for this class, Balcan, Hanneke, and Wortman [2008] claimed that for $\mathcal{P}$ the uniform distribution over $[0, 1]^k$, any $h \in \mathbb{C}$ with $\mathcal{P}(x : h(x) = +1) > 0$ has $\theta_h(0) < \infty$ (as a special case of their result for ordinary binary classification trees); as mentioned above, this is also implied by the later work of Friedman [2009], which further showed that $\limsup_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))/\varepsilon \lesssim k^{3/2}$; this can be refined to $\limsup_{\varepsilon \to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))/\varepsilon \lesssim k$ using a technique of Mahalanabis [2011, 2012]. Recent work by El-Yaniv and Wiener [2012], based on the general technique of El-Yaniv and Wiener [2010] in combination with the aforementioned analysis of Hanneke [2007a], shows that the label complexity of CAL for axis-aligned rectangles in the realizable case is $\lesssim \frac{k^3}{p}\mathrm{polylog}\left(\frac{k}{p\varepsilon\delta}\right)$, where $p = \mathcal{P}(x : f^\star(x) = +1) \vee \varepsilon$. Combined with the lower bound of Theorem 5.2, their proof also establishes that $\theta_h(\varepsilon) \lesssim \frac{k^3}{p}\mathrm{polylog}\left(\frac{k}{p\varepsilon}\right)$.

We should note that, in the realizable case, when $\mathcal{P}$ is a product distribution with a density, there is a simple active learning algorithm achieving label complexity $\lesssim \frac{1}{p\vee\varepsilon}\mathrm{Log}(1/\delta) + k\mathrm{Log}(k/\varepsilon)$, where $p = \mathcal{P}(x : f^\star(x) = +1)$. Consider the case of $\mathcal{P}$ uniform in $[0,1]^k$. The algorithm requests the labels $Y_1, Y_2, \ldots$ up until either exhausting the budget or finding the first t with $Y_t = +1$. In the former case, it returns $\hat{h} = \mathbb{1}^{\pm}_{\{\}}$, the always-negative classifier. In the latter case, it divides the remaining budget evenly to perform binary searches within a tiny-radius of the line extending parallel to each of the $2k$ coordinate directions and intersecting this positive point, in order to estimate the locations of the sides of the target rectangle; in the end, the algorithm returns the smallest rectangle consistent with the observed positive examples. If $p > \varepsilon$, and $n \gtrsim (1/p)\mathrm{Log}(1/\delta) + k\mathrm{Log}(k/\varepsilon)$, then with probability $1 - O(\delta)$, we will find a positive example within the first $(1/p)\mathrm{Log}(1/\delta)$ points in the sequence, and then the binary searches will identify the sides of the rectangles up to $\pm\varepsilon/k$ after at most $\lesssim k\mathrm{Log}(k/\varepsilon)$ additional queries, so that the total error rate is at most ε. On the other hand, if $p \leq \varepsilon$, then regardless of whether the algorithm encounters any positive examples or not, the smallest rectangle $\hat{h}$ consistent with the observed labels has $\mathrm{er}(\hat{h}) \leq \varepsilon$. The case of $\mathcal{P}$ a general product distribution with a density can be mapped to this uniform case by first rescaling the axes so that the distribution appears uniform, which does not change the hypothesis class [see Hanneke, 2007a]. The drawback of this algorithm is that, in order to perform the binary searches to locate the sides of the rectangles, we needed to focus the queries very close to a line running parallel to each axis, so that we can search for each face individually. To obtain enough unlabeled samples within these small regions, we would need to search through an enormous number of unlabeled samples from the

$\mathcal{Z}_{\mathbf{X}}$ sequence.

7.5 Realizing any Disagreement Coefficient Function

Interestingly, it is possible to construct scenarios ($\mathbb{C}$ and $\mathcal{P}$) for which the disagreement coefficients $\theta_h(\varepsilon)$ realize an essentially arbitrary dependence on ε for all $h \in \mathbb{C}$. This is formalized in the following theorem; the principal construction in the proof is from the work of Balcan, Hanneke, and Vaughan [2010].

Theorem 7.24. Let $g : (0,1) \to [1,\infty)$ be any nonincreasing function with $\varepsilon \mapsto \varepsilon g(\varepsilon)$ nondecreasing, and $\forall \varepsilon \in (0,1), g(\varepsilon) \leq 1/\varepsilon$. There exists a space $\mathcal{X}$, a distribution $\mathcal{P}$ over $\mathcal{X}$, and a hypothesis class $\mathbb{C}$ on $\mathcal{X}$ with $d = 1$ such that, $\forall h \in \mathbb{C}$, $\theta_h(\varepsilon) = O(g(\varepsilon))$ and $\theta_h(\varepsilon) \neq o(g(\varepsilon))$.

Proof. To prove Theorem 7.24, we first note that the case $g(\varepsilon) = O(1)$ is trivial, since we can take $\mathcal{X} = \{0\}$, and then $\mathbb{C} = \{h_y : y \in \mathcal{Y}\}$, where $h_y(0) = y$, has $\mathrm{vc}(\mathbb{C}) = 1$ and $\theta_h(\varepsilon) = 1$ for all $h \in \mathbb{C}$. Likewise, the case of $g(\varepsilon) = \Omega(1/\varepsilon)$ is also fairly trivial if we allow $\mathbb{C}$ to be uncountable; for instance, taking $\mathcal{X} = [0,1]$, $\mathcal{P}$ uniform on $\mathcal{X}$, and $\mathbb{C} = \{\mathbb{1}^{\pm}_{\{x\}} : x \in [0,1]\}$, every $h \in \mathbb{C}$ has $\mathrm{B}(h,0) = \mathbb{C}$, and hence $\mathrm{DIS}(\mathrm{B}(h,0)) = \mathcal{X}$, so that $\theta_h(\varepsilon) = 1/\varepsilon$ for all $\varepsilon > 0$. However, what is far more interesting (and challenging to prove) is that one can even construct a *countable* class $\mathbb{C}$ with this property (though with slightly larger VC dimension); we defer the description of such a construction to Section 7.6 below, which explores the behavior of $\theta_h(\varepsilon)$ for countable classes more broadly.

To address the remaining case of Theorem 7.24, suppose $g(\varepsilon) \neq \Omega(1/\varepsilon)$ and $g(\varepsilon) = \omega(1)$. In this case, even allowing an uncountable $\mathbb{C}$, the construction that realizes $g(\varepsilon)$ as a disagreement coefficient $\theta_h(\varepsilon)$ can be somewhat involved (though see Section 7.6 for a discussion of the countable case); Balcan, Hanneke, and Vaughan [2010] provide a construction for this case, motivated as follows. If we were only interested in satisfying $\theta_h(\varepsilon_0) \propto g(\varepsilon_0)$ for a given *fixed* value of $\varepsilon_0 > 0$, we could simply take $\mathcal{X} = \{0, 1, \ldots, \lceil g(\varepsilon_0) \rceil\}$, $\mathcal{P}$ with $\mathcal{P}(\{x\}) = \varepsilon_0/2$ for $x \in \{1, \ldots, \lceil g(\varepsilon_0) \rceil\}$, and $\mathbb{C} = \{\mathbb{1}^{\pm}_{\{x\}} : x \in \{1, \ldots, \lceil g(\varepsilon_0) \rceil\}\}$. This way, for any $h \in \mathbb{C}$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon_0)))/\varepsilon_0 = \mathcal{P}(\{1, \ldots, \lceil g(\varepsilon_0) \rceil\})/\varepsilon_0 = \lceil g(\varepsilon_0) \rceil/2$, so that $\theta_h(\varepsilon_0) = \lceil g(\varepsilon_0) \rceil / 2$. Of course, this construction would only work for that particular value of $\varepsilon = \varepsilon_0$, and we would in fact have $\theta_h(\varepsilon) = O(1)$. To correct for this, we merely need to repeat this construction for a whole sequence $\{\varepsilon_i\}_{i=1}^{\infty}$ of ε values having $\varepsilon_i \to 0$. That is, for each ε_i, there is some set of $\lceil g(\varepsilon_i) \rceil$ points, each having probability $\propto \varepsilon_i$, of which h labels exactly one as positive. As long as the ε_i sequence shrinks quickly enough (so that the sets of points corresponding to smaller ε_i values have small total probability), we are guaranteed $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon_i))) \propto \varepsilon_i \lceil g(\varepsilon_i) \rceil$ for each i. The only catch is that we need to be careful to also maintain $\mathrm{vc}(\mathbb{C}) = 1$; to achieve this, we will have many sets of $\lceil g(\varepsilon_i) \rceil$ points for each ε_i, and the particular set in which the positive point of h resides is uniquely determined by which points h labels positive among the sets of points corresponding to larger ε_j values; this can be concisely described as a kind of rooted tree structure on the points in $\mathcal{X}$, where the nodes at level i have probability $\propto \varepsilon_i$, and the points labeled positive by h form a path in the tree.

Now let us formalize this idea. Let $\{\ell_i\}_{i=1}^{\infty}$ be any sequence of strictly positive values with $\sum_{i=1}^{\infty} \ell_i = 1$ (e.g., $\ell_i = 2^{-i}$ would suffice). Let p_1 be any value in $(0, 1/2)$ with $g(p_1) \geq 4$ and $p_1 \lceil g(p_1) \rceil \leq \ell_1$, and let $t_1 = \lceil g(p_1) \rceil$. Now for each $i \in \mathbb{N} \setminus \{1\}$, inductively define p_i as any value in $(0, 1)$ such that $p_i \lceil g(p_i) \rceil \prod_{j=1}^{i-1} t_j \leq \ell_i$, $p_i \lceil g(p_i) \rceil \leq p_{i-1}/2$, and $g(p_i) \geq 2$, and let $t_i = \lceil g(p_i) \rceil$. We are guaranteed these values of p_i exist for all $i \in \mathbb{N}$, due to the facts that $g(\varepsilon) \neq \Omega(1/\varepsilon)$ (i.e., $\liminf_{\varepsilon \to 0} \varepsilon g(\varepsilon) = 0$) and $g(\varepsilon) = \omega(1)$. Finally, let $p_0 = 1 - \sum_{i=1}^{\infty} p_i \prod_{j=1}^{i} t_j$. Note that, since we have chosen each p_i so that $p_i \prod_{j=1}^{i} t_j \leq \ell_i$, we have $p_0 \geq 0$.

Now let $\mathcal{X}$ be a countable collection of distinct points $x_{\mathbf{z}}$, where $\mathbf{z} \in \{0\} \cup \{(z_1, \ldots, z_k) : k \in \mathbb{N}, \forall i \leq k, z_i \in \{1, \ldots, t_i\}\}$. We structure the space $\mathcal{X}$ as a rooted infinite tree, as follows. The element x_0 is the root node of the tree, and we define $\mathrm{Children}(x_0) = \{x_{(1)}, \ldots, x_{(t_1)}\}$. Then for any $k \in \mathbb{N}$ and $\mathbf{z} = (z_1, \ldots, z_k)$ s.t. $\forall i \leq k, z_i \in \{1, \ldots, t_i\}$, define $\mathrm{Children}(x_{\mathbf{z}}) = \{x_{(z_1, \ldots, z_k, j)} : j \in \{1, \ldots, t_{k+1}\}\}$. Also, for any $x \in \mathcal{X}$, inductively define $\mathrm{Subtree}(x) = \{x\} \cup \bigcup_{x' \in \mathrm{Children}(x)} \mathrm{Subtree}(x')$.

Now specify the probability measure $\mathcal{P}$ as follows. Define $\mathcal{P}(\{x_0\}) = p_0$. For any $k \in \mathbb{N}$, and $\mathbf{z} = (z_1, \ldots, z_k)$ with $\forall i \leq k, z_i \in \{1, \ldots, t_i\}$, define $\mathcal{P}(\{x_{\mathbf{z}}\}) = p_k$. In particular, note that $\mathcal{P}(\{x\}) \geq 0$ for all $x \in \mathcal{X}$, and $\mathcal{P}(\mathcal{X}) = p_0 + \sum_{k=1}^{\infty} p_k \prod_{i=1}^{k} t_i = 1$, so that this is a valid definition for $\mathcal{P}$. In terms of the tree structure, this assigns equal probability to all nodes at any given level of the tree.

Finally, we define the classifiers in $\mathbb{C}$ as follows. Let $Z = \{\{z_i\}_{i=1}^{\infty} : \forall i \in \mathbb{N}, z_i \in \{1, \ldots, t_i\}\}$. For every sequence $\mathbf{z} = \{z_i\}_{i=1}^{\infty} \in Z$, define the classifier $h_{\mathbf{z}}$ by the property that

$$\{x \in \mathcal{X} : h_{\mathbf{z}}(x) = +1\} = \{x_0\} \cup \left\{x_{(z_1,\ldots,z_k)} : k \in \mathbb{N}\right\}.$$

In terms of the tree, the set of nodes labeled $+1$ by $h_{\mathbf{z}}$ are precisely the nodes along an infinite path starting from the root and traversing the z_i^{th} branch from the node at level $i-1$ to reach level i, for each $i \in \mathbb{N}$. Then define $\mathbb{C} = \{h_{\mathbf{z}} : \mathbf{z} \in Z\}$. Note that, if $h \in \mathbb{C}$ and $x, x' \in \mathcal{X}$ have $h(x) = h(x') = +1$, then we must have $x \in \text{Subtree}(x')$ or $x' \in \text{Subtree}(x)$; supposing the latter (without loss of generality), any $h' \in \mathbb{C}$ with $h'(x) = -1$ cannot have $h'(x') = +1$. Therefore, $\text{vc}(\mathbb{C}) \leq 1$. Since $h_{(1,1,\ldots)}(x_{(1)}) = +1 \neq h_{(2,2,\ldots)}(x_{(1)})$, $\text{vc}(\mathbb{C}) \geq 1$ as well, so that the VC dimension is exactly 1.

Now fix any $\mathbf{z} = \{z_i\}_{i=1}^{\infty} \in Z$, and we will show that $\theta_{h_{\mathbf{z}}}(\varepsilon)$ is $O(g(\varepsilon))$ but not $o(g(\varepsilon))$. First, note that for any $\mathbf{z}' = \{z_i'\}_{i=1}^{\infty} \in Z$ with $\mathbf{z}' \neq \mathbf{z}$, letting $i_{\min}(\mathbf{z}, \mathbf{z}') = \min\{i \in \mathbb{N} : z_i' \neq z_i\}$, we have $\text{DIS}(\{h_{\mathbf{z}'}, h_{\mathbf{z}}\}) = \{x_{(z_1,\ldots,z_i)} : i \geq i_{\min}(\mathbf{z}, \mathbf{z}')\} \cup \{x_{(z_1',\ldots,z_i')} : i \geq i_{\min}(\mathbf{z}, \mathbf{z}')\}$, so that

$$\mathcal{P}(\text{DIS}(\{h_{\mathbf{z}'}, h_{\mathbf{z}}\})) = 2 \sum_{i \geq i_{\min}(\mathbf{z},\mathbf{z}')} p_i \in [2p_{i_{\min}(\mathbf{z},\mathbf{z}')}, 4p_{i_{\min}(\mathbf{z},\mathbf{z}')}], \quad (7.10)$$

where the upper bound follows from the fact that $p_i \leq p_{i-1}/2$ for all $i \geq 2$.

Fix any $\varepsilon \in (0, 2p_1)$, and let $\hat{i}_{\varepsilon} = \min\{i \in \mathbb{N} : \varepsilon \geq 2p_i\}$; note that $\hat{i}_{\varepsilon} \geq 2$. By (7.10),

$$\text{B}(h_{\mathbf{z}}, \varepsilon) \subseteq \{h_{\mathbf{z}'} : \mathbf{z}' \in Z, i_{\min}(\mathbf{z}, \mathbf{z}') \geq \hat{i}_{\varepsilon}\}.$$

In particular,

$$\text{DIS}(\text{B}(h_{\mathbf{z}}, \varepsilon)) \subseteq \text{Subtree}(x_{(z_1,\ldots,z_{\hat{i}_{\varepsilon}-1})}) \setminus \{x_{(z_1,\ldots,z_{\hat{i}_{\varepsilon}-1})}\}.$$

Since every $i \geq 2$ has $p_i \lceil g(p_i) \rceil \leq p_{i-1}/2$, by induction every $x \in \mathcal{X} \setminus \{x_0\}$ has $\mathcal{P}(\text{Subtree}(x)) \leq 2\mathcal{P}(\{x\})$. Applying this to each of the children of $x_{(z_1,\ldots,z_{\hat{i}_\varepsilon-1})}$, we have

$$\begin{aligned}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},\varepsilon))) &\leq 2\mathcal{P}(\mathrm{Children}(x_{(z_1,\ldots,z_{\hat{i}_\varepsilon-1})})) = 2p_{\hat{i}_\varepsilon} t_{\hat{i}_\varepsilon} \\ &= 2p_{\hat{i}_\varepsilon} \lceil g(p_{\hat{i}_\varepsilon}) \rceil \leq 4p_{\hat{i}_\varepsilon} g(p_{\hat{i}_\varepsilon}) \leq 4\varepsilon g(\varepsilon),\end{aligned}$$

where this last inequality is due to the assumed monotonicity of $\varepsilon \mapsto \varepsilon g(\varepsilon)$. Thus, we have $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},\varepsilon)))/\varepsilon \leq 4g(\varepsilon) = O(g(\varepsilon))$, and therefore Lemma 7.9 implies $\theta_{h_{\mathbf{z}}}(\varepsilon) = O(g(\varepsilon))$.

For the remaining claim of $\theta_{h_{\mathbf{z}}}(\varepsilon) \neq o(g(\varepsilon))$, note that for any $k \in \mathbb{N}$, by (7.10),

$$\mathrm{B}(h_{\mathbf{z}}, 4p_k) \supseteq \{h_{\mathbf{z}'} : \mathbf{z}' \in Z, i_{\min}(\mathbf{z}, \mathbf{z}') \geq k\},$$

so that in particular,

$$\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, 4p_k)) \supseteq \mathrm{Children}(x_{(z_1,\ldots,z_{k-1})}).$$

Therefore,

$$\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, 4p_k))) \geq \mathcal{P}(\mathrm{Children}(x_{(z_1,\ldots,z_{k-1})}) = p_k t_k \geq p_k g(p_k).$$

Since $p_k \to 0$ as $k \to \infty$, we find that $\limsup_{\varepsilon\to 0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},4\varepsilon)))/\varepsilon}{g(\varepsilon)} \geq 1$. Therefore, since $g(\varepsilon) \geq g(4\varepsilon)$, we have $\limsup_{\varepsilon\to 0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},\varepsilon)))/\varepsilon}{g(\varepsilon)} = \limsup_{\varepsilon\to 0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},4\varepsilon)))/(4\varepsilon)}{g(4\varepsilon)} \geq (1/4)\limsup_{\varepsilon\to 0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},4\varepsilon)))/\varepsilon}{g(\varepsilon)} \geq 1/4$, which implies $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}},\varepsilon)))/\varepsilon \neq o(g(\varepsilon))$. Together with Lemma 7.9, this implies $\theta_{h_{\mathbf{z}}}(\varepsilon) \neq o(g(\varepsilon))$ as well. It is interesting to note that the proof of this $\theta_{h_{\mathbf{z}}}(\varepsilon) \neq o(g(\varepsilon))$ half of the theorem did not require monotonicity of $\varepsilon \mapsto \varepsilon g(\varepsilon)$. □

The constructions in the proof of Theorem 7.24 can be implemented by a variety of commonly-used hypothesis classes $\mathbb{C}$, including k-dimensional linear separators (for any $k \geq 2$) and axis-aligned rectangles in $\mathbb{R}^k$ (again, for $k \geq 2$), among others. In fact, the scenarios constructed for the case $g(\varepsilon) \neq \Omega(1/\varepsilon)$ can even be implemented by these classes using distributions $\mathcal{P}$ having densities (with respect to λ^k), so that Theorem 7.15 cannot generally be improved without additional assumptions.

Balcan, Hanneke, and Vaughan [2010] also prove that, for the scenarios constructed for the case $g(\varepsilon) \neq O(1)$ and $g(\varepsilon) \neq \Omega(1/\varepsilon)$ in the above proof, for any active learning algorithm, there exists a distribution $\mathcal{P}_{XY}$ in the realizable case (with marginal distribution $\mathcal{P}$ over $\mathcal{X}$), such that the algorithm's label complexity Λ has $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \neq o(g(\varepsilon))$; this is essentially the best asymptotic lower bound one can achieve for these scenarios, since there is a simple algorithm achieving $O(g(\varepsilon)\log(1/\varepsilon))$ label complexity for every such $\mathcal{P}_{XY}$ in the realizable case: namely, the algorithm that queries the nodes at depth one until identifying a positive point, and then recurses on the subtree rooted at that node. In fact, this algorithm achieves label complexity $O(g(\varepsilon))$ if we require each $i \geq 2$ to satisfy $g(p_i) \geq cg(p_{i-1})$ for some constant $c > 1$, which is possible due to $g(\varepsilon) \neq O(1)$.

7.6 Countable Classes

One of the insights in the work of Balcan, Hanneke, and Vaughan [2010] is that, if the distribution $\mathcal{P}$ is known to us at the time of designing the algorithm, we can often get around trivial negative results by running the algorithms with $\tilde{\mathbb{C}}$ a *countable dense subset* of $\mathbb{C}$, in place of $\mathbb{C}$: that is, a countable subset $\tilde{\mathbb{C}} \subseteq \mathbb{C}$ that satisfies $\sup_{h \in \mathbb{C}} \inf_{g \in \tilde{\mathbb{C}}} \mathcal{P}(x : g(x) \neq h(x)) = 0$. Such a set $\tilde{\mathbb{C}}$ is always guaranteed to exist; for instance, if $d < \infty$, we can take $\tilde{\mathbb{C}} = \bigcup_i \mathbb{C}_i$, where $\mathbb{C}_i$ is a minimum-sized subset of $\mathbb{C}$ with $\bigcup_{h \in \mathbb{C}_i} \mathrm{B}(h, 2^{-i}) = \mathbb{C}$; this minimum size (i.e., the 2^{-i}-covering number) is guaranteed to be finite by a result of Haussler [1995]. It is not hard to see that Theorems 5.1 and 5.4 continue to hold if we replace the initialization "$V \leftarrow \mathbb{C}$" in CAL and RobustCAL with "$V \leftarrow \tilde{\mathbb{C}}$", while using the disagreement coefficient with respect to $\tilde{\mathbb{C}}$ in the label complexity bound, rather than that with respect to $\mathbb{C}$. In some cases, this improves the results, since the resulting disagreement coefficient can sometimes be smaller (e.g., consider $\mathbb{C} = \{\mathbb{1}^{\pm}_{\{x\}} : x \in [0,1]\}$ and $\mathcal{P}$ uniform over $\mathcal{X} = [0,1]$). As such, it becomes interesting for us to study the behavior of the disagreement coefficient with respect to countable classes. Countable classes additionally offer us a variety of conveniences that are not necessarily available for uncountable classes. This section describes several known results regarding the behavior of the disagreement coefficients for countable classes. Some of these properties played important roles in the analysis of Balcan, Hanneke, and Vaughan [2010], while others are interesting in themselves or as important examples. I should note that this section is somewhat more technical than the others in this article, and can be skipped by the casual reader without significant loss of continuity.

As a first observation, we note that there is an interesting analogy between having $\mathcal{P}(\partial h) = 0$ and having an implication that convergence in probability implies convergence almost surely. This analogy can be formalized as in the following lemma (here we denote by "$\xrightarrow{p}$" convergence in probability, and by "$\xrightarrow{a.s.}$" convergence almost surely).

Lemma 7.25. Let $X \sim \mathcal{P}$. If $\mathbb{C}$ is countable, then for any classifier h, the following two statements are equivalent.

- $\mathcal{P}(\partial h) = 0$.
- For every sequence $\{h_n\}_{n=1}^{\infty}$ in $\mathbb{C}$, $h_n(X) \xrightarrow{p} h(X) \implies h_n(X) \xrightarrow{a.s.} h(X)$.

Furthermore, if $\mathcal{P}(\partial h) > 0$, there exists a sequence $\{g_n\}_{n=1}^{\infty}$ in $\mathbb{C}$ with $\forall n, m \in \mathbb{N}, n \neq m \Rightarrow \mathcal{P}(x : g_n(x) \neq g_m(x)) > 0$, such that $g_n(X) \xrightarrow{p} h(X)$ and $\mathcal{P}(\{x : g_n(x) \nrightarrow h(x)\} \oplus \partial h) = 0$, where $\oplus$ denotes the symmetric difference.

Proof. We prove these equivalences in the contrapositive. Suppose $\mathcal{P}(\partial h) > 0$. For each $\varepsilon > 0$, since $\mathrm{B}(h, \varepsilon)$ is countable, continuity of probability measures implies there exists a *finite* subset $\mathcal{H}_\varepsilon \subseteq \mathrm{B}(h, \varepsilon)$ with $\mathcal{P}(\mathrm{DIS}(\mathcal{H}_\varepsilon)) \geq (1-\varepsilon)\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, \varepsilon)))$. Enumerate the elements of $\mathcal{H}_\varepsilon$ as a finite sequence $\{g_{\varepsilon,1}, \ldots, g_{\varepsilon,|\mathcal{H}_\varepsilon|}\}$. Now form a sequence $\{h_n\}_{n=1}^{\infty}$ by concatenating these finite sequences for decreasing values of $\varepsilon \in \{1/(m+1) : m \in \mathbb{N}\}$: that is, $h_n = g_{\varepsilon_n, k_n}$, where $\varepsilon_n = \min\{1/(m+1) : m \in \mathbb{N}, n > \sum_{\ell < m} |\mathcal{H}_{1/(\ell+1)}|\}$ and $k_n = n - \sum_{\ell < m} |\mathcal{H}_{1/(\ell+1)}|$. Note that $\mathcal{P}(x : h_n(x) \neq h(x)) \leq \varepsilon_n \to 0$, so that $h_n(X) \xrightarrow{p} h(X)$. However, any $x \in \lim_{n\to\infty} \mathrm{DIS}(\{h_m : m \geq n\})$ necessarily has $h_n(x) \nrightarrow h(x)$, so that $\mathcal{P}(x : h_n(x) \nrightarrow h(x)) \geq \mathcal{P}\left(\lim_{n\to\infty} \mathrm{DIS}(\{h_m : m \geq n\})\right)$; by continuity of probability measures, $\mathcal{P}\left(\lim_{n\to\infty} \mathrm{DIS}(\{h_m : m \geq n\})\right) = \lim_{n\to\infty} \mathcal{P}(\mathrm{DIS}(\{h_m : m \geq n\}))$; by monotonicity, this equals $\lim_{n\to\infty} \mathcal{P}(\mathrm{DIS}(\mathcal{H}_{\varepsilon_n})) \geq \lim_{n\to\infty}(1 -$

$\varepsilon_n)\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,\varepsilon_n))) \geq \lim_{n\to\infty}(1-\varepsilon_n)\mathcal{P}(\partial h) = \mathcal{P}(\partial h) > 0$. In total, we have $\mathcal{P}(x : h_n(x) \nrightarrow h(x)) = \mathcal{P}(\partial h) > 0$, which disproves almost sure convergence of $h_n(X)$ to $h(X)$.

For the other direction, suppose $\{h_n\}_{n=1}^{\infty}$ is some sequence in $\mathbb{C}$ with $h_n(X) \xrightarrow{p} h(X)$, but for which almost sure convergence of $h_n(X)$ to $h(X)$ fails to hold. Let $\mathcal{X}_0$ denote the set of $x \in \mathcal{X}$ for which $h_n(x)$ converges, but does not converge to $h(x)$: that is, $\mathcal{X}_0 = \liminf_{n\to\infty}\{x : h_n(x) \neq h(x)\}$. Since $h_n(X) \xrightarrow{p} h(X)$ implies $\mathcal{P}(x : h_n(x) \neq h(x)) \to 0$, and monotonicity and continuity of probability measures imply $\lim_{n\to\infty}\mathcal{P}(x : h_n(x) \neq h(x)) \geq \lim_{n\to\infty}\mathcal{P}\left(\bigcap_{m\geq n}\{x : h_m(x) \neq h(x)\}\right) = \mathcal{P}\left(\liminf_{n\to\infty}\{x : h_n(x) \neq h(x)\}\right) = \mathcal{P}(\mathcal{X}_0)$, we have $\mathcal{P}(\mathcal{X}_0) = 0$. Thus, since almost sure convergence of $h_n(X)$ to $h(X)$ does not hold, it must be that the set $\mathcal{X}_1$ of $x \in \mathcal{X}$ for which $h_n(x)$ does not converge satisfies $\mathcal{P}(\mathcal{X}_1) > 0$. Note that $\mathcal{X}_1 = \lim_{n\to\infty}\mathrm{DIS}(\{h_m : m \geq n\})$. Furthermore, for any $\varepsilon > 0$, since $\mathcal{P}(x : h_n(x) \neq h(x)) \to 0$, any sufficiently large $m \in \mathbb{N}$ satisfies $h_m \in \mathrm{B}(h,\varepsilon)$; therefore, any $x \in \mathcal{X}_1$ has $x \in \mathrm{DIS}(\mathrm{B}(h,\varepsilon))$ as well. Since this is true for all $\varepsilon > 0$, we have $\mathcal{X}_1 \subseteq \bigcap_{\varepsilon>0}\mathrm{DIS}(\mathrm{B}(h,\varepsilon)) = \partial h$. Thus, $\mathcal{P}(\partial h) \geq \mathcal{P}(\mathcal{X}_1) > 0$.

It remains to establish the final claim. Let $\{h_n\}_{n=1}^{\infty}$ be the sequence constructed in the first paragraph of this proof. Since $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,0))) \leq \sum_{g\in\mathrm{B}(h,0)}\mathcal{P}(x : g(x) \neq h(x)) = 0$, and we have shown above that $\lim_{n\to\infty}\mathcal{P}\left(\mathrm{DIS}(\{h_m : m \geq n\})\right) > 0$, there must be an infinite subsequence of values $n \in \mathbb{N}$ with $h_n \notin \mathrm{B}(h,0)$. A union bound implies that, for any $g \in \mathbb{C}$, any $n \in \mathbb{N}$ with $\mathcal{P}(x : h_n(x) \neq g(x)) = 0$ has $\mathcal{P}(x : h_n(x) \neq h(x)) \geq \mathcal{P}(x : g(x) \neq h(x)) - \mathcal{P}(x : h_n(x) \neq g(x)) = \mathcal{P}(x : g(x) \neq h(x))$; since $\mathcal{P}(x : h_n(x) \neq h(x)) \to 0$, this implies that for any $g \in \mathbb{C}\setminus\mathrm{B}(h,0)$, there are at most a finite number of $n \in \mathbb{N}$ with $\mathcal{P}(x : h_n(x) \neq g(x)) = 0$. In particular, this implies there is an infinite subsequence of indices $n \in \mathbb{N}$ with $h_n \notin \mathrm{B}(h,0)$ and $\forall m > n, \mathcal{P}(x : h_n(x) \neq h_m(x)) > 0$; denote by $\{n_k\}_{k=1}^{\infty}$ the sequence of all such indices (in increasing order). Now for any $n \in \mathbb{N}$, and any $x_0 \in \mathrm{DIS}(\{h_m : m \geq n\})$, either $x_0 \in \mathrm{DIS}(\mathrm{B}(h,0) \cup \{h\}) \cup \bigcap_{m\geq n : h_m \notin \mathrm{B}(h,0)}\mathrm{DIS}(\{h_m, h\})$, or else there exist $\ell, m \geq n$ with $h_\ell, h_m \notin \mathrm{B}(h,0)$ and $h_\ell(x_0) \neq h_m(x_0)$. In this latter case, by the above argument, $\exists j, k \in \mathbb{N}$ s.t. $n_j \geq \ell, n_k \geq m$, and $\mathcal{P}(x : h_\ell(x) \neq h_{n_j}(x)) = \mathcal{P}(x : h_m(x) \neq h_{n_k}(x)) = 0$. For these indices j, k, either $h_{n_j}(x_0) \neq h_{n_k}(x_0)$, $h_{n_j}(x_0) \neq h_\ell(x_0)$, or $h_{n_k}(x_0) \neq h_m(x_0)$; in any case, $x_0 \in \mathrm{DIS}(\{h_{n_i} : n_i \geq n\}) \cup \mathrm{DIS}(\mathrm{B}(h_\ell, 0)) \cup \mathrm{DIS}(\mathrm{B}(h_m, 0))$. In total, letting $\mathcal{X}_2 = \bigcup_{m=1}^{\infty} \mathrm{DIS}(\mathrm{B}(h_m, 0))$, and $\mathcal{X}_3 = \lim_{\ell\to\infty} \bigcap_{m\geq\ell : h_m \notin \mathrm{B}(h,0)} \mathrm{DIS}(\{h_m, h\})$ and noting that $\bigcap_{m\geq n : h_m\notin\mathrm{B}(h,0)} \mathrm{DIS}(\{h_m, h\}) \subseteq \mathcal{X}_3$, we have $\mathrm{DIS}(\{h_m : m \geq n\}) \subseteq \mathrm{DIS}(\{h_{n_i} : n_i \geq n\}) \cup \mathcal{X}_2 \cup \mathcal{X}_3 \cup \mathrm{DIS}(\mathrm{B}(h, 0) \cup \{h\})$, which implies $\mathrm{DIS}(\{h_{n_i} : n_i \geq n\}) \supseteq \mathrm{DIS}(\{h_m : m \geq n\}) \setminus (\mathcal{X}_2 \cup \mathcal{X}_3 \cup \mathrm{DIS}(\mathrm{B}(h, 0) \cup \{h\}))$. Therefore, monotonicity and the union bound imply $\mathcal{P}\left(\lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) = \mathcal{P}\left(\lim_{n\to\infty} \mathrm{DIS}(\{h_{n_i} : n_i \geq n\})\right) \geq \mathcal{P}\left(\lim_{n\to\infty} \mathrm{DIS}(\{h_m : m \geq n\})\right) - \mathcal{P}(\mathcal{X}_2) - \mathcal{P}(\mathcal{X}_3) - \mathcal{P}(DIS(\mathrm{B}(h, 0) \cup \{h\}))$. A union bound implies $\mathcal{P}(\mathcal{X}_2) \leq \sum_{m=1}^{\infty} \sum_{g\in\mathrm{B}(h_m,0)} \mathcal{P}(x : g(x) \neq h_m(x)) = 0$, and similarly $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, 0) \cup \{h\})) \leq \sum_{g\in\mathrm{B}(h,0)} \mathcal{P}(x : g(x) \neq h(x)) = 0$. Also, by continuity of probability measures, $\mathcal{P}(\mathcal{X}_3) = \lim_{\ell\to\infty} \mathcal{P}\left(\bigcap_{m\geq\ell : h_m\notin\mathrm{B}(h,0)} \mathrm{DIS}(\{h_m, h\})\right) \leq \lim_{\ell\to\infty} \sup_{m\geq\ell} \mathcal{P}(\mathrm{DIS}(\{h_m, h\})) = 0$. Since we have already established above that $\mathcal{P}\left(\lim_{n\to\infty} \mathrm{DIS}(\{h_m : m \geq n\})\right) \geq \mathcal{P}(\partial h)$, we have $\mathcal{P}\left(\lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\}\right) \geq \mathcal{P}(\partial h)$. Since $\{h_{n_i} : i \geq k\} \subseteq \mathrm{B}(h, \sup_{i\geq k} \mathcal{P}(x : h_{n_i}(x) \neq h(x)))$ and $\mathcal{P}(x : h_{n_i}(x) \neq h(x)) \to 0$, we find that $\lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\}) \subseteq \partial h$. It therefore holds that $\mathcal{P}\left(\partial h \oplus \lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) = \mathcal{P}(\partial h) - \mathcal{P}\left(\lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) = 0$. Finally, since $\{x : h_{n_k}(x) \not\to h(x)\} = \left(\lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) \cup \left(\liminf_{k\to\infty}\{x : h_{n_k}(x) \neq h(x)\}\right)$, a union bound implies $\mathcal{P}\left(\{x : h_{n_k}(x) \not\to h(x)\} \oplus \partial h\right) \leq \mathcal{P}\left(\{x : h_{n_k}(x) \not\to h(x)\} \oplus \lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) + \mathcal{P}\left(\partial h \oplus \lim_{k\to\infty} \mathrm{DIS}(\{h_{n_i} : i \geq k\})\right) = \mathcal{P}\left(\liminf_{k\to\infty}\{x : h_{n_k}(x) \neq h(x)\}\right) \leq \liminf_{k\to\infty} \mathcal{P}(x : h_{n_k}(x) \neq h(x)) = 0$. Thus, the claim holds by taking $g_k = h_{n_k}$ for each $k \in \mathbb{N}$. $\square$

The following lemma describes a property that will be important in discussions of sequences converging in probability but not converging almost surely.

Lemma 7.26. For any sequence $\{A_n\}_{n=1}^{\infty}$ of measurable subsets of $\mathcal{X}$ with $\lim_{n\to\infty} \mathcal{P}(A_n) = 0$, and any measurable set $C \subseteq \limsup_{n\to\infty} A_n$ with $\mathcal{P}(C) > 0$, $\forall \varepsilon > 0, \exists n \in \mathbb{N}$ s.t. $0 < \mathcal{P}(A_n \cap C) \leq \mathcal{P}(A_n) < \varepsilon$.

Proof. Fix any $\varepsilon > 0$; since $\mathcal{P}(A_n) \to 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $\sup_{n \geq N_\varepsilon} \mathcal{P}(A_n) < \varepsilon$. Since $\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \subseteq \bigcup_{n=N_\varepsilon}^{\infty} A_n$, monotonicity and a union bound imply $\mathcal{P}(C) = \mathcal{P}\left(C \cap \limsup_{n\to\infty} A_n\right) \leq \mathcal{P}\left(C \cap \bigcup_{n=N_\varepsilon}^{\infty} A_n\right) = \mathcal{P}\left(\bigcup_{n=N_\varepsilon}^{\infty} (A_n \cap C)\right) \leq \sum_{n=N_\varepsilon}^{\infty} \mathcal{P}(A_n \cap C)$; since $\mathcal{P}(C) > 0$, there must be some $n \geq N_\varepsilon$ with $\mathcal{P}(A_n \cap C) > 0$. Furthermore, by monotonicity and the definition of N_ε, $\mathcal{P}(A_n \cap C) \leq \mathcal{P}(A_n) < \varepsilon$. $\square$

The above lemmas lead to another basic observation: namely, that countable classes of VC dimension 1 have $\mathcal{P}(\partial h) = 0$ for all classifiers h (including those not in $\mathbb{C}$); in particular, Lemma 7.12 implies $\theta_h(\varepsilon) = o(1/\varepsilon)$ for these classes. The proof essentially shows that, if $\mathcal{P}(\partial h) > 0$, then there must be such a diverse set of classifiers in $\mathbb{C}$ that we can shatter at least 2 points in ∂h. Specifically, we have the following theorem.

Theorem 7.27. If $\mathbb{C}$ is countable and $d = 1$, then every classifier h and distribution $\mathcal{P}$ satisfy $\theta_h(\varepsilon) = o(1/\varepsilon)$.

Proof. Fix any classifier h and distribution $\mathcal{P}$, and suppose $\mathbb{C}$ is countable with $d = 1$. Let $X \sim \mathcal{P}$. Lemma 7.12 and Lemma 7.25 imply it suffices to show that every sequence $\{h_n\}_{n=1}^{\infty}$ in $\mathbb{C}$ with $h_n(X) \xrightarrow{p} h(X)$ also has $h_n(X) \xrightarrow{a.s.} h(X)$. If there is no sequence $\{h_n\}_{n=1}^{\infty}$ in $\mathbb{C}$ with $h_n(X) \xrightarrow{p} h(X)$, this is trivially satisfied; otherwise, suppose $\{h_n\}_{n=1}^{\infty}$ is a sequence in $\mathbb{C}$ satisfying $h_n(X) \xrightarrow{p} h(X)$.

For the purpose of contradiction, suppose almost sure convergence of $h_n(X)$ to $h(X)$ fails to hold: that is, the set $A = \{x : h_n(x) \nrightarrow h(x)\}$ has $\mathcal{P}(A) > 0$. Letting $A_n = \{x : h_n(x) \neq h(x)\}$, we have $A = \limsup_{n\to\infty} A_n$, and furthermore $h_n(X) \xrightarrow{p} h(X)$ implies $\mathcal{P}(A_n) \to 0$. Lemma 7.26 implies $\exists n_1 \in \mathbb{N}$ with $\mathcal{P}(A_{n_1} \cap A) > 0$. In light of this, Lemma 7.26 (with $C = A_{n_1} \cap A$ and $\varepsilon = \mathcal{P}(A_{n_1} \cap A)$) further implies $\exists n_2 \in \mathbb{N}$ with $0 < \mathcal{P}(A_{n_2} \cap A_{n_1} \cap A) < \mathcal{P}(A_{n_1} \cap A)$. In particular, this also implies $\mathcal{P}(A_{n_1} \cap A_{n_2}^c \cap A) = \mathcal{P}(A_{n_1} \cap A) - \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A) > 0$ (where

$A_{n_2}^c = \mathcal{X} \setminus A_{n_2}$). Applying Lemma 7.26 a final time (this time, with $C = A_{n_1} \cap A_{n_2}^c \cap A$ and $\varepsilon = \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A)$), we find $\exists n_3 \in \mathbb{N}$ with $0 < \mathcal{P}(A_{n_3} \cap A_{n_1} \cap A_{n_2}^c \cap A) \leq \mathcal{P}(A_{n_3}) < \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A)$. This also implies that $\mathcal{P}(A_{n_1} \cap A_{n_2} \cap A_{n_3}^c \cap A) = \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A) - \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A_{n_3} \cap A) \geq \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A) - \mathcal{P}(A_{n_3}) > 0$. Finally, since $\mathcal{P}(A_n) \to 0$, $\exists n_4 \in \mathbb{N}$ with $\mathcal{P}(A_{n_4}) < \min \{\mathcal{P}(A_{n_1} \cap A_{n_2}^c \cap A_{n_3} \cap A), \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A_{n_3}^c \cap A)\}$, so that $\mathcal{P}(A_{n_1} \cap A_{n_2}^c \cap A_{n_3} \cap A_{n_4}^c \cap A) \geq \mathcal{P}(A_{n_1} \cap A_{n_2}^c \cap A_{n_3} \cap A) - \mathcal{P}(A_{n_4}) > 0$ and likewise $\mathcal{P}(A_{n_1} \cap A_{n_2} \cap A_{n_3}^c \cap A_{n_4}^c \cap A) \geq \mathcal{P}(A_{n_1} \cap A_{n_2} \cap A_{n_3}^c \cap A) - \mathcal{P}(A_{n_4}) > 0$. In particular, this implies there exist $x_1 \in A_{n_1} \cap A_{n_2}^c \cap A_{n_3} \cap A_{n_4}^c$ and $x_2 \in A_{n_1} \cap A_{n_2} \cap A_{n_3}^c \cap A_{n_4}^c$. Reflecting on the definitions of these sets, we see that each of $(h_{n_1}(x_1), h_{n_1}(x_2)), (h_{n_2}(x_1), h_{n_2}(x_2)), (h_{n_3}(x_1), h_{n_3}(x_2))$, and $(h_{n_4}(x_1), h_{n_4}(x_2))$ represent distinct classifications of the pair (x_1, x_2); specifically, $x_1, x_2 \in A_{n_4}^c$ implies h_{n_4} classifies these points in agreement with h, while $x_1, x_2 \in A_{n_1}$ implies h_{n_1} classifies them opposite h, and similarly we see that h_{n_2} classifies x_1 in agreement with h while it classifies x_2 opposite h, and h_{n_3} classifies x_1 opposite h and classifies x_2 in agreement with h. But since $\{h_{n_1}, h_{n_2}, h_{n_3}, h_{n_4}\} \subseteq \mathbb{C}$, this implies that $\mathbb{C}$ shatters (x_1, x_2), which contradicts $d = 1$. □

Note that this is essentially the best one can show at this level of generality, since the construction in the proof of the $g(\varepsilon) \neq \Omega(1/\varepsilon)$ case of Theorem 7.24 (which has $\mathrm{vc}(\mathbb{C}) = 1$) can easily be modified so that $\mathbb{C}$ is a countable class: for instance, in the context of that proof, the subset of classifiers $\{h_{\mathbf{z}} : \mathbf{z} = (z_1, z_2, \ldots) \in Z, \lim_{i \to \infty} z_i = 1\}$, which is in fact a countable dense subset of the original class $\mathbb{C}$, would also suffice by essentially the same argument. Additionally, note that, in light of the discussion above regarding the use of countable dense subsets in CAL and RobustCAL, Theorem 7.27 also has positive implications for the label complexity of active learning with uncountable classes $\mathbb{C}$ of $\mathrm{vc}(\mathbb{C}) = 1$.

As alluded to in the proof of Theorem 7.24, results such as Theorem 7.27 are not generally available for classes with higher VC dimension. Indeed, it is possible to construct scenarios in which $\mathbb{C}$ is countable and has VC dimension 2, yet $\mathcal{P}$ is constructed so that *every* classifier $h \in \mathbb{C}$ has $\theta_h(\varepsilon) = \Omega(1/\varepsilon)$. The idea behind this construction is to build the classifiers in $\mathbb{C}$ in the style of *interval* classifiers. This is natural, since we know the empty interval classifier has $1/\varepsilon$ disagreement coefficient with respect to the class of intervals under any nonatomic distribution; this remains true even if we restrict the class to those intervals that have rational-valued boundary points, which is a countable set of classifiers. However, the class of rational intervals itself is not sufficient for this result, since it includes classifiers with nonzero probability in their positive region, which therefore have finite constant bounds on their respective disagreement coefficients. We thus need to somehow modify the classifiers so that every classifier is an empty interval in some region of the space, while at the same time maintaining that empty intervals have disagreement coefficient $\Omega(1/\varepsilon)$. We formalize this construction in the proof of the following theorem.

Theorem 7.28. There exists a space $\mathcal{X}$, a distribution $\mathcal{P}$ over $\mathcal{X}$, and a countable hypothesis class $\mathbb{C}$ with $d = 2$, such that every $h \in \mathbb{C}$ has $\theta_h(\varepsilon) = \Omega(1/\varepsilon)$.

Proof. Consider $\mathcal{X} = (0, 1)$ and $\mathcal{P}$ uniform over $\mathcal{X}$. We specify the classifiers in the hypothesis class $\mathbb{C}$ as follows. Noting that the set $\mathcal{I} = \bigcup_{k=1}^{\infty} \mathbb{N}^k$ is countable, let $J : \mathcal{I} \to \mathbb{N}$ be a bijection between $\mathcal{I}$ and $\mathbb{N}$. Furthermore, let $R = (R_1, R_2)$ be a bijection mapping $\mathbb{N}$ to the set of ordered pairs $\{(p, q) \in \mathbb{Q} \times \mathbb{Q} : 0 \leq p < q < 1\}$, where we write $R(i) = (R_1(i), R_2(i))$; this being a countable set, such a bijection must exist. For each integer $k > 1$ and each $i_1, \ldots, i_k \in \mathbb{N}$, define the classifier $h_{(i_1,\ldots,i_k)}$ by the property that, $\forall x \in \mathcal{X}$, $h_{(i_1,\ldots,i_k)}(x) = +1$ if and only if

$$x \in \bigcup_{j=1}^{k-1} [2^{-J(i_1,\ldots,i_j)}(1 + R_1(i_{j+1})), 2^{-J(i_1,\ldots,i_j)}(1 + R_2(i_{j+1}))).$$

Thus, $h_{(i_1,\ldots,i_k)}$ corresponds to a carefully-constructed union of rational intervals. Finally, define the class

$$\mathbb{C} = \{h_{(i_1,\ldots,i_k)} : k \in \mathbb{N} \setminus \{1\}, i_1, \ldots, i_k \in \mathbb{N}\}.$$

The classifiers $h_{(i_1,\ldots,i_k)}$ have the following interpretation. The space $(0, 1)$ is partitioned into a countably infinite number of disjoint nonempty subintervals; specifically, for each $k \in \mathbb{N}$ and $i_1, \ldots, i_k \in \mathbb{N}$, define the region $N_{J(i_1,\ldots,i_k)} = [2^{-J(i_1,\ldots,i_k)}, 2^{1-J(i_1,\ldots,i_k)})$. Furthermore, for $k \in \mathbb{N} \setminus \{1\}$ and $i_1, \ldots, i_k \in \mathbb{N}$, the classifier $h_{(i_1,\ldots,i_k)}$ labels positive a certain nonempty rational interval in each of the regions $N_{J(i_1)}, \ldots, N_{J(i_1,\ldots,i_{k-1})}$; it labels all other points negative, and in particular, labels all of $N_{J(i_1,\ldots,i_k)}$ as negative. For any $i_{k+1} \in \mathbb{N}$, the classifier $h_{(i_1,\ldots,i_k,i_{k+1})}$ agrees with $h_{(i_1,\ldots,i_k)}$ on the labels of all points outside the region $N_{J(i_1,\ldots,i_k)}$, but has a non-empty rational interval in $N_{J(i_1,\ldots,i_k)}$ (whose end-points are uniquely indexed by the value i_{k+1}). In fact, the set $\{h_{(i_1,\ldots,i_k,i_{k+1})} : i_{k+1} \in \mathbb{N}\}$ realizes every possible rational interval $[p, q)$ with $p < q$ and $p, q \in N_{J(i_1,\ldots,i_k)}$, while agreeing with $h_{(i_1,\ldots,i_k)}$ on all other points. Thus, within the region $N_{J(i_1,\ldots,i_k)}$, the classifier $h_{(i_1,\ldots,i_k)}$ represents the empty interval in a rational-intervals subclass.

For any integer $k > 1$ and $i_1, \ldots, i_k \in \mathbb{N}$, for any $r > 0$, the set of rational intervals $\{[p, q) : p, q \in \mathbb{Q} : 0 \le p < q < 1, (p - q) \le r/2^{-J(i_1,\ldots,i_k)}\}$ covers the space $[0, 1)$; this set can be expressed equivalently as $\{[R_1(i_{k+1}), R_2(i_{k+1})) : i_{k+1} \in \mathbb{N}, (R_2(i_{k+1}) - R_1(i_{k+1}))2^{-J(i_1,\ldots,i_k)} \le r\}$. This implies the set of intervals $\{[2^{-J(i_1,\ldots,i_k)}R_1(i_{k+1}), 2^{-J(i_1,\ldots,i_k)}R_2(i_{k+1})) : i_{k+1} \in \mathbb{N}, (R_2(i_{k+1}) - R_1(i_{k+1}))2^{-J(i_1,\ldots,i_k)} \le r\}$ covers the space $[0, 2^{-J(i_1,\ldots,i_k)})$; shifting this by $2^{-J(i_1,\ldots,i_k)}$, we have that $\bigcup\{[2^{-J(i_1,\ldots,i_k)}(1 + R_1(i_{k+1})), 2^{-J(i_1,\ldots,i_k)}(1 + R_2(i_{k+1}))) : i_{k+1} \in \mathbb{N}, (R_2(i_{k+1}) - R_1(i_{k+1}))2^{-J(i_1,\ldots,i_k)} \le r\} = N_{J(i_1,\ldots,i_k)}$. Since every $i_{k+1} \in \mathbb{N}$ has $\mathrm{DIS}(\{h_{(i_1,\ldots,i_k,i_{k+1})}, h_{(i_1,\ldots,i_k)}\}) = [2^{-J(i_1,\ldots,i_k)}(1 + R_1(i_{k+1})), 2^{-J(i_1,\ldots,i_k)}(1 + R_2(i_{k+1})))$, we have that

$$
\begin{aligned}
&\mathrm{DIS}(\mathrm{B}(h_{(i_1,\ldots,i_k)}, r)) \\
&\supseteq \bigcup \Big\{ \mathrm{DIS}(\{h_{(i_1,\ldots,i_k,i_{k+1})}, h_{(i_1,\ldots,i_k)}\}) : \\
&\qquad\qquad i_{k+1} \in \mathbb{N}, \mathcal{P}(\mathrm{DIS}(\{h_{(i_1,\ldots,i_k,i_{k+1})}, h_{(i_1,\ldots,i_k)}\})) \le r \Big\} \\
&= \bigcup \left\{ \left[2^{-J(i_1,\ldots,i_k)}(1 + R_1(i_{k+1})), 2^{-J(i_1,\ldots,i_k)}(1 + R_2(i_{k+1}))\right) : \right. \\
&\qquad\qquad \left. i_{k+1} \in \mathbb{N}, (R_2(i_{k+1}) - R_1(i_{k+1}))2^{-J(i_1,\ldots,i_k)} \le r \right\} \\
&= N_{J(i_1,\ldots,i_k)},
\end{aligned}
$$

so that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{(i_1,\ldots,i_k)}, r))) \ge \mathcal{P}(N_{J(i_1,\ldots,i_k)}) = 2^{-J(i_1,\ldots,i_k)} > 0$, and hence $\forall \varepsilon > 0$, $\theta_{h_{(i_1,\ldots,i_k)}}(\varepsilon) = 1 \vee \sup_{r>\varepsilon} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{(i_1,\ldots,i_k)}, r)))/r \ge 2^{-J(i_1,\ldots,i_k)}/\varepsilon = \Omega(1/\varepsilon)$.

Finally, to show $\mathrm{vc}(\mathbb{C}) = 2$, since Theorem 7.27 implies $\mathrm{vc}(\mathbb{C}) > 1$, it suffices to show that no 3 points are shattered by $\mathbb{C}$. Let $x_1, x_2, x_3 \in (0,1)$, and without loss of generality, suppose $x_1 \le x_2 \le x_3$. Now suppose $\mathbb{C}$ shatters (x_1, x_3). There exists some $k \in \mathbb{N}$ and $i_1, \ldots, i_k \in \mathbb{N}$ such that $x_1 \in N_{J(i_1,\ldots,i_k)}$, and some $k' \in \mathbb{N}$ and $i'_1, \ldots, i'_{k'} \in \mathbb{N}$ such that $x_3 \in N_{J(i'_1,\ldots,i'_{k'})}$. Since (x_1, x_3) is shattered by $\mathbb{C}$, there exists $k'' \in \mathbb{N} \setminus \{1\}$ and $i''_1, \ldots, i''_{k''} \in \mathbb{N}$ such that $x_1, x_3 \in \bigcup_{j=1}^{k''-1} [2^{-J(i''_1,\ldots,i''_j)}(1 + R_1(i_{j+1})), 2^{-J(i''_1,\ldots,i''_j)}(1 + R_2(i_{j+1}))) \subseteq \bigcup_{j=1}^{k''-1} N_{J(i''_1,\ldots,i''_j)}$. Since the N_j sets are disjoint, we must have that $\forall j \le k, i_j = i''_j$ and $\forall j \le k', i'_j = i''_j$. If $k' < k$, then any $k''' \in \mathbb{N} \setminus \{1\}$ and $i'''_1, \ldots, i'''_{k'''} \in \mathbb{N}$ with $h_{(i'''_1,\ldots,i'''_{k'''})}(x_3) = -1$ must have either $k''' \le k'$ or some $j \le k' + 1$ with $i'''_j \ne i''_j = i_j$; since $k' + 1 \le k$, either of these cases would imply $h_{(i'''_1,\ldots,i'''_{k'''})}$ classifies all of $N_{J(i_1,\ldots,i_k)}$ as -1, and in particular, $h_{(i'''_1,\ldots,i'''_{k'''})}(x_1) = -1$. Similarly, if $k < k'$, any $h \in \mathbb{C}$ with $h(x_1) = -1$ must have $h(x_3) = -1$ as well. Therefore, since $\mathbb{C}$ shatters (x_1, x_3), we have $k = k'$, which implies $(i_1, \ldots, i_k) = (i'_1, \ldots, i'_{k'})$, and thus $N_{J(i_1,\ldots,i_k)} = N_{J(i'_1,\ldots,i'_{k'})}$. In particular, this implies that any classifier $h \in \mathbb{C}$ with $h(x_1) = h(x_3) = +1$ has $\{x : h(x) = +1\} \supseteq [x_1, x_3]$; since $x_2 \in [x_1, x_3]$, this implies $h(x_2) = +1$ as well, so that $\mathbb{C}$ does not shatter (x_1, x_2, x_3). $\square$

Although Theorem 7.28 indicates that we are not always guaranteed $o(1/\varepsilon)$ disagreement coefficients for countable classes of VC dimension 2 (as we *are* for countable classes of VC dimension 1), there are interesting weaker positive results one can show. For instance, if $\mathcal{H}$ is a (possibly uncountable) set of classifiers with $\mathrm{vc}(\mathcal{H}) = 2$, with no two distinct $h, g \in \mathcal{H}$ having $\mathcal{P}(x : h(x) \neq g(x)) = 0$, and $\mathbb{C}$ is countable with $\mathbb{C} \subseteq \mathcal{H}$, then one can show that the set of classifiers $h \in \mathcal{H}$ with $\theta_h(\varepsilon) \neq o(1/\varepsilon)$ is at most countably infinite. This result is not necessarily available for uncountable sets $\mathcal{H}$ with $\mathrm{vc}(\mathcal{H}) > 2$, by a straightforward modification of the proof of Theorem 7.28.

Another type of positive result for countable classes involves the behavior of $\mathcal{P}(\partial h)$ as a function of h. Interestingly, we can show that as long as $\mathbb{C}$ is countable and $d < \infty$, *every* classifiers h with $\inf_{g\in\mathbb{C}} \mathcal{P}(x : h(x) \neq g(x)) = 0$ has $\inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g) = 0$ for all $r > 0$. One implication of this is that, although Theorem 7.28 implies that (when $d \geq 2$) it is possible for every $h \in \mathbb{C}$ to have $\theta_h(\varepsilon) \geq c_h + c'_h/\varepsilon = \Omega(1/\varepsilon)$, for some h-dependent constants $c_h, c'_h > 0$, we are guaranteed that the set of classifiers h in $\mathbb{C}$ with very *small* c'_h is *dense* in the space $\mathbb{C}$. The reasoning behind this result is that, if the classifiers g close to h all have $\mathcal{P}(\partial g)$ bounded away from zero, then some of their disagreement cores must overlap; in fact, there will be entire pockets of these classifiers having almost-identical disagreement cores, and we can then find shatterable sets in these overlap regions, using carefully-chosen classifiers responsible for these disagreements; furthermore, these classifiers that witness the shatterability can also be chosen from within these pockets where the disagreement cores are all similar, so that their disagreement cores will also overlap, and hence we can repeat the argument, adding more points to the shatterable sets; this allows us to inductively construct shatterable sets of arbitrary finite sizes, thereby obtaining a contradiction to the fact that $d < \infty$. This argument is formalized in the following lemma and theorem.

Lemma 7.29. Let $\mathbb{C}^*$ denote the set of all classifiers. For any classifier h, $\lim_{\varepsilon\to 0} \mathcal{P}\left(\bigcup_{g\in\mathrm{B}_{\mathbb{C}^*}(h,\varepsilon)} \partial g \setminus \partial h\right) = 0$.

Proof. Fix any $\varepsilon > 0$. Since every $g \in \mathrm{B}_{\mathbb{C}^*}(h,\varepsilon)$ has $\mathrm{B}(g,\varepsilon) \subseteq \mathrm{B}(h,2\varepsilon)$, and therefore $\mathrm{DIS}(\mathrm{B}(g,\varepsilon)) \subseteq \mathrm{DIS}(\mathrm{B}(h,2\varepsilon))$, we have $\bigcup_{g\in\mathrm{B}_{\mathbb{C}^*}(h,\varepsilon)} \partial g \subseteq \bigcup_{g\in\mathrm{B}_{\mathbb{C}^*}(h,\varepsilon)} \mathrm{DIS}(\mathrm{B}(g,\varepsilon)) \subseteq \mathrm{DIS}(\mathrm{B}(h,2\varepsilon))$. Therefore, $\mathcal{P}\left(\bigcup_{g\in\mathrm{B}_{\mathbb{C}^*}(h,\varepsilon)} \partial g \setminus \partial h\right) \leq \mathcal{P}\left(\mathrm{DIS}(\mathrm{B}(h,2\varepsilon)) \setminus \partial h\right) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,2\varepsilon))) - \mathcal{P}(\partial h)$. The claim now follows by taking the limit as $\varepsilon \to 0$, since continuity of probability measures implies $\lim_{\varepsilon\to 0} \mathcal{P}(\mathrm{DIS}(\mathrm{B}(h,2\varepsilon))) = \mathcal{P}\left(\lim_{\varepsilon\to 0} \mathrm{DIS}(\mathrm{B}(h,2\varepsilon))\right)$ and $\lim_{\varepsilon\to 0} \mathrm{DIS}(\mathrm{B}(h,2\varepsilon)) = \partial h$ by definition of ∂h. □

Lemma 7.29 essentially says that every g sufficiently close to h necessarily has ∂g almost-entirely contained in ∂h. With this lemma in hand, we have the following theorem.

Theorem 7.30. If $\mathbb{C}$ is countable and $d < \infty$, then any classifier h with $\inf_{g\in\mathbb{C}} \mathcal{P}(x : g(x) \neq h(x)) = 0$ has $\sup_{r>0} \inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g) = 0$.

Proof. First note that, since $\inf_{g\in\mathbb{C}} \mathcal{P}(x : g(x) \neq h(x)) = 0$, we know $\mathrm{B}(h,r) \neq \emptyset$ for all $r > 0$, so that $\inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g)$ is well-defined.

We prove this result by contradiction; for this purpose, suppose $\sup_{r>0} \inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g) > 0$. Since $r \mapsto \inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g)$ is bounded, nonincreasing, and bounded away from zero (by assumption), for any $\delta \in (0,1)$, there exists $r_\delta \in (0,1)$ with $\inf_{g\in\mathrm{B}(h,r_\delta)} \mathcal{P}(\partial g) > \sup_{r>0} \inf_{g\in\mathrm{B}(h,r)} \mathcal{P}(\partial g)/(1+\delta) > 0$; in particular, letting $q_\delta = \inf_{g\in\mathrm{B}(h,r_\delta)} \mathcal{P}(\partial g)$, we have $q_\delta \leq \inf_{g\in\mathrm{B}(h,r_\delta/2)} \mathcal{P}(\partial g) < (1+\delta)q_\delta$. Now for any $\delta \in (0,1)$, let $f_\delta \in \mathrm{B}(h, r_\delta/2)$ be a classifier with $\mathcal{P}(\partial f_\delta) < (1+\delta)q_\delta$, and let $\varepsilon_\delta \in (0, r_\delta/2)$ be any sufficiently small value so that every $g \in \mathrm{B}(f_\delta, \varepsilon_\delta)$ has $\mathcal{P}(\partial g \setminus \partial f_\delta) < \mathcal{P}(\partial f_\delta)\delta^2/(1+\delta)$; such a value ε_δ is guaranteed to exist by Lemma 7.29. We will prove by induction that, for any $n \in \mathbb{N}$, $\delta \in (0, 2^{1-n})$, and $\varepsilon \in (0, \varepsilon_\delta)$, there exists a set of classifiers $\mathcal{H}_n \subseteq \mathrm{B}(f_\delta, \varepsilon)$ with $|\mathcal{H}_n| = 2^n$ and n sets $A_1, \ldots, A_n \subseteq \mathcal{X}$ with $\min_{i\in\{1,\ldots,n\}} \mathcal{P}(A_i) > 0$ such that $\mathcal{H}_n$ shatters every $(x_1, \ldots, x_n) \in A_1 \times \cdots \times A_n$.

As a base case, for $n = 1$, for any $\delta \in (0,1)$ and $\varepsilon \in (0, \varepsilon_\delta)$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f_\delta, \varepsilon))) \geq \mathcal{P}(\partial f_\delta) > 0$ while $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f_\delta, 0))) \leq \sum_{g\in\mathrm{B}(f_\delta,0)} \mathcal{P}(x : g(x) \neq f_\delta(x)) = 0$, so there must exist some $g \in \mathrm{B}(f_\delta, \varepsilon) \setminus \mathrm{B}(f_\delta, 0)$; since $\mathcal{P}(x : f_\delta(x) \neq g(x)) > 0$ for any such g, we can satisfy the claim by taking $\mathcal{H}_1 = \{f_\delta, g\}$ and $A_1 = \{x : f_\delta(x) \neq g(x)\}$.

Now for the inductive step, suppose this claim is satisfied with $n = m - 1$ for some $m \in \mathbb{N} \setminus \{1\}$, and we will show this implies the claim also holds with $n = m$. Fix any $\delta \in (0, 2^{1-m})$ and $\varepsilon \in (0, \varepsilon_\delta)$, and consider a set $\mathcal{H}_{m-1} \subseteq \mathrm{B}(f_\delta, \varepsilon/2)$ with $|\mathcal{H}_{m-1}| = 2^{m-1}$ and sets $A_1^{m-1}, \ldots, A_{m-1}^{m-1} \subseteq \mathcal{X}$ with $\min_{i \leq m-1} \mathcal{P}(A_i^{m-1}) > 0$; these are guaranteed to exist by the inductive hypothesis. Since $\varepsilon/2 < \varepsilon_\delta < r_\delta/2$ and $f_\delta \in \mathrm{B}(h, r_\delta/2)$, we have $\mathrm{B}(f_\delta, \varepsilon/2) \subseteq \mathrm{B}(h, r_\delta)$, so any $g \in \mathrm{B}(f_\delta, \varepsilon/2)$ also has $\mathcal{P}(\partial g) \geq q_\delta > \mathcal{P}(\partial f_\delta)/(1+\delta)$; in particular, by definition of ε_δ, this implies $\mathcal{P}(\partial f_\delta \cap \partial g) = \mathcal{P}(\partial g) - \mathcal{P}(\partial g \setminus \partial f_\delta) > \mathcal{P}(\partial f_\delta)\left(\frac{1}{1+\delta} - \frac{\delta^2}{1+\delta}\right) = (1-\delta)\mathcal{P}(\partial f_\delta)$, so that $\mathcal{P}(\partial f_\delta \setminus \partial g) = \mathcal{P}(\partial f_\delta) - \mathcal{P}(\partial f_\delta \cap \partial g) < \mathcal{P}(\partial f_\delta) - (1-\delta)\mathcal{P}(\partial f_\delta) = \delta \mathcal{P}(\partial f_\delta)$. Therefore, $\mathcal{P}\left(\bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) \geq \mathcal{P}\left(\partial f_\delta \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) = \mathcal{P}(\partial f_\delta) - \mathcal{P}\left(\partial f_\delta \cap \bigcup_{g \in \mathcal{H}_{m-1}} (\partial g)^c\right) \geq \mathbb{P}(\partial f_\delta) - \sum_{g \in \mathcal{H}_{m-1}} \mathcal{P}(\partial f_\delta \setminus \partial g) > \mathcal{P}(\partial f_\delta) - |\mathcal{H}_{m-1}| \delta \mathcal{P}(\partial f_\delta) = \mathcal{P}(\partial f_\delta)(1 - 2^{m-1}\delta) > 0$, where the first inequality is by monotonicity, the second inequality is by a union bound, the third inequality was proven above, and the final inequality follows from the fact that $\delta \in (0, 2^{1-m})$.

Now enumerate the classifiers in $\mathcal{H}_{m-1} = \{g_1, \ldots, g_{2^{m-1}}\}$, and let $\gamma \in \left(0, \min\left\{\varepsilon/2, 2^{1-m} \min_{i \leq m-1} \mathcal{P}(A_i^{m-1})\right\}\right)$. For the purpose of (nested) induction, suppose that, for some $k \in \{1, \ldots, 2^{m-1}\}$, there exist classifiers $g_i' \in \mathrm{B}(g_i, \gamma)$ for each $i \in \{1, \ldots, k-1\}$, such that, letting $G_{k-1} = \bigcap_{i=1}^{k-1} \{x : g_i'(x) \neq g_i(x)\}$ (or $G_{k-1} = \mathcal{X}$ if $k = 1$), we have $\mathcal{P}\left(G_{k-1} \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$. Since $\mathcal{P}(\partial g_k) > 0$, Lemma 7.25 implies there exists a sequence $\{h_t\}_{t=1}^\infty$ in $\mathbb{C}$ with $\mathcal{P}(x : h_t(x) \neq g_k(x)) \to 0$ and $\mathcal{P}\left((\limsup_{t \to \infty} \{x : h_t(x) \neq g_k(x)\}) \oplus \partial g_k\right) = 0$; in particular, this implies $C = (\limsup_{t \to \infty} \{x : h_t(x) \neq g_k(x)\}) \cap G_{k-1} \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g$ has $\mathcal{P}(C) = \mathcal{P}\left(G_{k-1} \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$; therefore, by Lemma 7.26, $\exists t \in \mathbb{N}$ such that $\mathcal{P}(x : h_t(x) \neq g_k(x)) < \gamma$ and $\mathcal{P}(\{x : h_t(x) \neq g_k(x)\} \cap C) > 0$: that is, if we define $g_k' = h_t$ for this t, and $G_k = \bigcap_{i=1}^k \{x : g_i'(x) \neq g_i(x)\}$, we have $g_k' \in \mathrm{B}(g_k, \gamma)$ and $\mathcal{P}\left(G_k \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$, thus completing the inductive step. Since we have defined $G_0 = \mathcal{X}$, we have already proven above that $\mathcal{P}\left(G_0 \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$, and therefore, by the principle of induction, there exist classifiers $g_i' \in \mathrm{B}(g_i, \gamma)$, for each $i \in \{1, \ldots, 2^{m-1}\}$, such that the set $G_{2^{m-1}} = \bigcap_{i=1}^{2^{m-1}} \{x : g_i'(x) \neq g_i(x)\}$

satisfies $\mathcal{P}\left(G_{2^{m-1}} \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$.

Returning to the broader inductive argument, we are now ready to define the sets A_i. Specifically, for each $i \in \{1, \ldots, m-1\}$, define $A_i = A_i^{m-1} \cap \bigcap_{k=1}^{2^{m-1}} \{x : g_k(x) = g'_k(x)\}$, and further define $A_m = \bigcap_{k=1}^{2^{m-1}} \{x : g_k(x) \neq g'_k(x)\} = G_{2^{m-1}}$. We have proven that $\mathcal{P}(A_m) \geq \mathcal{P}\left(G_{2^{m-1}} \cap \bigcap_{g \in \mathcal{H}_{m-1}} \partial g\right) > 0$. Furthermore, since $g'_k \in \mathrm{B}(g_k, \gamma)$ for each $k \in \{1, \ldots, 2^{m-1}\}$, monotonicity and a union bound imply that every $i \in \{1, \ldots, m-1\}$ has $\mathcal{P}(A_i) = \mathcal{P}(A_i^{m-1}) - \mathcal{P}\left(A_i^{m-1} \cap \bigcup_{k=1}^{2^{m-1}} \{x : g_k(x) \neq g'_k(x)\}\right) \geq \mathcal{P}(A_i^{m-1}) - \sum_{k=1}^{2^{m-1}} \mathcal{P}(x : g_k(x) \neq g'_k(x)) \geq \mathcal{P}(A_i^{m-1}) - 2^{m-1}\gamma > 0$, where this last inequality follows from the fact that $\gamma < 2^{1-m}\mathcal{P}(A_i^{m-1})$. Finally, define $\mathcal{H}_m = \mathcal{H}_{m-1} \cup \{g'_k : k \in \{1, \ldots, 2^{m-1}\}\}$, and let $(x_1, \ldots, x_m) \in A_1 \times \cdots \times A_m$. Fix any values $y_1, \ldots, y_m \in \mathcal{Y}$. The inductive hypothesis implies $\mathcal{H}_{m-1}$ shatters $(x_1, \ldots, x_{m-1})$, so there exists some $k \in \{1, \ldots, 2^{m-1}\}$ for which $g_k(x_i) = y_i$ for all $i \leq m-1$. Furthermore, for each $i \leq m-1$, by definition of A_i, we have $g'_k(x_i) = g_k(x_i) = y_i$ as well. However, by definition of A_m, we have $g_k(x_m) \neq g'_k(x_m)$, so that (since $\mathcal{Y}$ has only two elements) either $g_k(x_m) = y_m$ or $g'_k(x_m) = y_m$; in either case, $\exists g \in \mathcal{H}_m$ such that, $\forall i \leq m, g(x_i) = y_i$. Since $y_1, \ldots, y_m$ were arbitrary, we conclude that $\mathcal{H}_m$ shatters $(x_1, \ldots, x_m)$, which completes the inductive proof of the theorem. □

Balcan, Hanneke, and Vaughan [2010] employed a slightly-more-sophisticated variant of the above argument to show that, even if $\mathbb{C}$ is uncountable, as long as $d < \infty$, it is possible to decompose $\mathbb{C}$ into a countable collection of subsets $\mathbb{C}_i$, each of which has $\mathcal{P}(\partial_{\tilde{\mathbb{C}}_i} h) = 0$ for every $h \in \mathbb{C}_i$, where $\tilde{\mathbb{C}}_i$ is a countable dense subset of $\mathbb{C}_i$. They used the existence of this decomposition, along with a variant of CAL and a model selection algorithm for active learning (a variant of the ActiveSelect procedure of Section 8.6 below), to prove the surprisingly general result that, for any distribution $\mathcal{P}$, and any hypothesis class $\mathbb{C}$ (possibly uncountable) with $d < \infty$, there exists an active learning algorithm achieving a label complexity $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) = o(1/\varepsilon)$ for every $\mathcal{P}_{XY}$ in the realizable case having marginal $\mathcal{P}$ over $\mathcal{X}$. We discuss this result, and a more general variant of it, in Section 8.6. We conclude

Chapter 7 with a formal construction of this decomposition, due to Balcan, Hanneke, and Vaughan [2010]; note that $\mathbb{C}$ is *not* required to be countable for this theorem.

Theorem 7.31. If $d < \infty$, then for any distribution $\mathcal{P}$, there exists a sequence $\{\mathbb{C}_i\}_{i=1}^{\infty}$ of disjoint subsets of $\mathbb{C}$ with $\bigcup_{i=1}^{\infty} \mathbb{C}_i = \mathbb{C}$ such that, for every $i \in \mathbb{N}$, and any countable dense subset $\tilde{\mathbb{C}}_i$ of $\mathbb{C}_i$, every $h \in \mathbb{C}_i$ has $\mathcal{P}(\partial_{\tilde{\mathbb{C}}_i} h) = 0$.

Before going through the proof of this result, we first have two minor lemmas.

Lemma 7.32. For any set $\mathcal{H}$ of classifiers, and any countable dense subsets $\tilde{\mathcal{H}}_1, \tilde{\mathcal{H}}_2 \subseteq \mathcal{H}$, every classifier f satisfies $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}_1} f\right) = \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}_2} f\right)$.

Proof. Fix any $\varepsilon > 0$, and enumerate the classifiers in $\mathrm{B}_{\tilde{\mathcal{H}}_1}(f,\varepsilon) = \{h_n : n \in \mathbb{N}\}$. For each $n \in \mathbb{N}$, let $g_n \in \tilde{\mathcal{H}}_2$ be such that $\mathcal{P}(x : g_n(x) \neq h_n(x)) < 2^{-n}\varepsilon$; note that $\mathcal{P}(x : g_n(x) \neq f(x)) \leq \mathcal{P}(x : g_n(x) \neq h_n(x)) + \mathcal{P}(x : h_n(x) \neq f(x)) < 2\varepsilon$, so that $g_n \in \mathrm{B}_{\tilde{\mathcal{H}}_2}(f, 2\varepsilon)$. Therefore, $\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_1}(f,\varepsilon)) = \bigcup_{n,m\in\mathbb{N}} \mathrm{DIS}(\{h_n,h_m\}) \subseteq \bigcup_{n,m\in\mathbb{N}} \mathrm{DIS}(\{g_n,g_m\}) \cup \mathrm{DIS}(\{g_n,h_n\}) \cup \mathrm{DIS}(\{g_m,h_m\}) \subseteq \mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_2}(f,2\varepsilon)) \cup \bigcup_{n\in\mathbb{N}} \mathrm{DIS}(\{g_n,h_n\})$. By a union bound, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_1}(f,\varepsilon))) \leq \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_2}(f,2\varepsilon))) + \sum_{n\in\mathbb{N}} \mathcal{P}(x : g_n(x) \neq h_n(x)) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_2}(f,2\varepsilon))) + \varepsilon$. By continuity of probability measures, we therefore have $\mathcal{P}(\partial_{\tilde{\mathcal{H}}_1} f) = \mathcal{P}\left(\lim_{\varepsilon\to 0} \mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_1}(f,\varepsilon))\right) = \lim_{\varepsilon\to 0} \mathcal{P}\left(\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_1}(f,\varepsilon))\right) \leq \lim_{\varepsilon\to 0} \mathcal{P}\left(\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_2}(f,2\varepsilon))\right) + \varepsilon = \mathcal{P}\left(\lim_{\varepsilon\to 0} \mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}_2}(f,2\varepsilon))\right) = \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}_2} f\right)$. By symmetry, the same argument implies $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}_2} f\right) \leq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}_1} f\right)$, and combined these imply the theorem. $\square$

Lemma 7.33. For any set $\mathcal{H}$ of classifiers, and any nonempty $\mathcal{G} \subseteq \mathcal{H}$, letting $\tilde{\mathcal{H}}$ and $\tilde{\mathcal{G}}$ be countable dense subsets of $\mathcal{H}$ and $\mathcal{G}$, respectively, every classifier h satisfies $\mathcal{P}\left(\partial_{\tilde{\mathcal{G}}} h\right) \leq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}} h\right)$.

Proof. Since every $\varepsilon > 0$ has $\mathrm{B}_{\tilde{\mathcal{G}}}(h,\varepsilon) \subseteq \mathrm{B}_{\tilde{\mathcal{H}}\cup\tilde{\mathcal{G}}}(h,\varepsilon)$, and hence $\mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{G}}}(h,\varepsilon)) \subseteq \mathrm{DIS}(\mathrm{B}_{\tilde{\mathcal{H}}\cup\tilde{\mathcal{G}}}(h,\varepsilon))$, we have $\partial_{\tilde{\mathcal{G}}} h \subseteq \partial_{\tilde{\mathcal{H}}\cup\tilde{\mathcal{G}}} h$, so that

$\mathcal{P}\left(\partial_{\tilde{\mathcal{G}}} h\right) \leq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}} \cup \tilde{\mathcal{G}}} h\right)$. Furthermore, since $\tilde{\mathcal{H}} \cup \tilde{\mathcal{G}}$ is also a countable dense subset of $\mathcal{H}$, Lemma 7.32 implies $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}} \cup \tilde{\mathcal{G}}} h\right) = \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}} h\right)$. □

We are now ready to prove the theorem.

Proof of Theorem 7.31. Let $\mathcal{H}^0_{()} = \mathbb{C}$, and let $\tilde{\mathcal{H}}^0_{()}$ be a countable dense subset of $\mathbb{C}$. For each $t \in \{1, \ldots, d\}$, and each $i_1, \ldots, i_t \in \mathbb{N} \cup \{0\}$, inductively define the following sets. If $i_t \neq 0$, define

$$\mathcal{H}^t_{(i_1,\ldots,i_t)} = \Bigg\{ h \in \mathcal{H}^{t-1}_{(i_1,\ldots,i_{t-1})} : \left(1 + 2^{-d}\right)^{-i_t} < \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{t-1}_{(i_1,\ldots,i_{t-1})}} h\right) \leq \left(1 + 2^{-d}\right)^{1-i_t} \Bigg\},$$

and if $i_t = 0$, define

$$\mathcal{H}^t_{(i_1,\ldots,i_t)} = \left\{ h \in \mathcal{H}^{t-1}_{(i_1,\ldots,i_{t-1})} : \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{t-1}_{(i_1,\ldots,i_{t-1})}} h\right) = 0 \right\}.$$

Also, in either case, let $\tilde{\mathcal{H}}^t_{(i_1,\ldots,i_t)}$ denote a countable dense subset of $\mathcal{H}^t_{(i_1,\ldots,i_t)}$. Since $\{0\} \cup \bigcup_{i \in \mathbb{N}} \left(\left(1 + 2^{-d}\right)^{-i}, \left(1 + 2^{-d}\right)^{1-i} \right] = [0, 1]$, we have $\bigcup_{i_t \in \mathbb{N} \cup \{0\}} \mathcal{H}^t_{(i_1,\ldots,i_t)} = \mathcal{H}^{t-1}_{(i_1,\ldots,i_{t-1})}$; furthermore, since these $\left(\left(1 + 2^{-d}\right)^{-i}, \left(1 + 2^{-d}\right)^{1-i} \right]$ intervals are disjoint, and do not contain 0, the sets $\{\mathcal{H}^t_{(i_1,\ldots,i_t)} : i_t \in \mathbb{N} \cup \{0\}\}$ are disjoint as well. Therefore, by induction, this defines a collection of disjoint sets $\{\mathcal{H}^d_{(i_1,\ldots,i_d)} : i_1, \ldots, i_d \in \mathbb{N} \cup \{0\}\}$ such that $\bigcup_{i_1,\ldots,i_d \in \mathbb{N} \cup \{0\}} \mathcal{H}^d_{(i_1,\ldots,i_d)} = \mathcal{H}^0_{()} = \mathbb{C}$. Since $(\mathbb{N} \cup \{0\})^d$ is countable, this collection of disjoint subsets of $\mathbb{C}$ is also countable, and therefore can be enumerated as a sequence $\{\mathbb{C}_i\}_{i=1}^{\infty}$; in particular, we will take this as the definition of the sequence in the theorem statement, so that it only remains to prove that, for any indices $i_1, \ldots, i_d \in \mathbb{N} \cup \{0\}$, every $h \in \mathcal{H}^d_{(i_1,\ldots,i_d)}$ has $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^d_{(i_1,\ldots,i_d)}} h\right) = 0$.

For the sake of a proof by contradiction, suppose some $i_1, \ldots, i_d \in \mathbb{N} \cup \{0\}$ and $h \in \mathcal{H}^d_{(i_1,\ldots,i_d)}$ have $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^d_{(i_1,\ldots,i_d)}} h\right) > 0$. We will prove by induction that this fact would enable us to shatter $d + 1$ points, thus contradicting $\mathrm{vc}(\mathbb{C}) = d$. First, note that, by Lemma 7.33, any such $i_1, \ldots, i_d$ must have $i_j \neq 0$ for each $j \leq d$, since $\mathcal{H}^d_{(i_1,\ldots,i_d)} \subseteq \mathcal{H}^j_{(i_1,\ldots,i_j)}$. For brevity, define $\mathcal{H}^k = \mathcal{H}^k_{(i_1,\ldots,i_k)}$ and $\tilde{\mathcal{H}}^k = \tilde{\mathcal{H}}^k_{(i_1,\ldots,i_k)}$ for each $k \in \{0, \ldots, d\}$. For each $j \in \{1, \ldots, d\}$, let $\varepsilon_j > 0$ be any sufficiently small value so that every $g \in \mathrm{B}_{\mathcal{H}^j}(h, \varepsilon_j)$ has $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{j-1}} g \setminus \partial_{\tilde{\mathcal{H}}^{j-1}} h\right) < \mathcal{P}(\partial_{\tilde{\mathcal{H}}^{j-1}} h) 4^{-d}/(1 + 2^{-d})$; such a value ε_j must exist by Lemma 7.29. Let $\varepsilon_{\min} = \min_{j \leq d} \varepsilon_j$. We will prove by induction that, for any $n \in \{1, \ldots, d+1\}$, and $\varepsilon \in (0, \varepsilon_{\min})$, there exists a set of classifiers $\mathcal{G}^n \subseteq \mathrm{B}_{\mathcal{H}^{d+1-n}}(h, \varepsilon)$ with $|\mathcal{G}^n| = 2^n$, and n sets $A_1, \ldots, A_n \subseteq \mathcal{X}$ with $\min_{i \in \{1,\ldots,n\}} \mathcal{P}(A_i) > 0$ such that $\mathcal{G}^n$ shatters every $(x_1, \ldots, x_n) \in A_1 \times \cdots \times A_n$.

As a base case, if $n = 1$, for any $\varepsilon > 0$, since $\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}_{\tilde{\mathcal{H}}^d}(h, \varepsilon)\right)\right) \geq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^d} h\right) > 0$, and $\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}_{\tilde{\mathcal{H}}^d}(h, 0)\right)\right) \leq \sum_{g \in \mathrm{B}_{\tilde{\mathcal{H}}^d}(h,0)} \mathcal{P}(x : g(x) \neq h(x)) = 0$, there must exist a classifier $g \in \mathrm{B}_{\tilde{\mathcal{H}}^d}(h, \varepsilon) \setminus \mathrm{B}_{\tilde{\mathcal{H}}^d}(h, 0)$; since any such g has $\mathcal{P}(x : g(x) \neq h(x)) > 0$, we can satisfy the claim by taking $\mathcal{G}^1 = \{h, g\}$ and $A_1 = \{x : g(x) \neq h(x)\}$.

Now take as an inductive hypothesis that the claim is satisfied for $n = m - 1$ for some $m \in \{2, \ldots, d+1\}$. Fix any $\varepsilon \in (0, \varepsilon_{\min})$, and consider a set $\mathcal{G}^{m-1} \subseteq \mathrm{B}_{\mathcal{H}^{d+1-(m-1)}}(h, \varepsilon/2)$ with $|\mathcal{G}^{m-1}| = 2^{m-1}$ and sets $A_1^{m-1}, \ldots, A_{m-1}^{m-1} \subseteq \mathcal{X}$ with $\min_{i \leq m-1} \mathcal{P}(A_i^{m-1}) > 0$ such that every $(x_1, \ldots, x_{m-1}) \in A_1^{m-1} \times \cdots \times A_{m-1}^{m-1}$ is shattered by $\mathcal{G}^{m-1}$; these are guaranteed to exist by the inductive hypothesis. Since $\varepsilon/2 < \varepsilon_{\min}$, each $g \in \mathcal{G}^{m-1}$ satisfies $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} g \setminus \partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right) < \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right) 4^{-d}/(1 + 2^{-d})$. Additionally, since $\mathcal{G}^{m-1} \subseteq \mathcal{H}^{d+1-(m-1)}$, each $g \in \mathcal{G}^{m-1}$ satisfies $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > \left(1 + 2^{-d}\right)^{-i_{d+1-(m-1)}}$; furthermore, $h \in \mathcal{H}^{d+1-(m-1)}$, so that $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right) \leq \left(1 + 2^{-d}\right)^{1-i_{d+1-(m-1)}}$ as well. Combining the above, we have $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right) / \left(1 + 2^{-d}\right)$, so that $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} h \cap \partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) = \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) - \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} g \setminus \partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right) > \left(\frac{1-4^{-d}}{1+2^{-d}}\right) \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}} h\right)$, and therefore

$$\begin{aligned}\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h \setminus \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) &= \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) - \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h \cap \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) \\ &< \left(1 - \frac{1-4^{-d}}{1+2^{-d}}\right)\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) = 2^{-d}\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right).\end{aligned}$$

By a union bound, we also have

$$\begin{aligned}\mathcal{P}\left(\bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) &\geq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h \cap \bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) \\ &= \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) - \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h \cap \bigcup_{g\in\mathcal{G}^{m-1}} \left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}\right)^c\right) \\ &\geq \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) - \sum_{g\in\mathcal{G}^{m-1}} \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h \setminus \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right).\end{aligned}$$

Combined with the above, this implies

$$\mathcal{P}\left(\bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) > \mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) - |\mathcal{G}^{m-1}|2^{-d}\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}h\right) \geq 0,$$

where this last inequality follows from the fact that $|\mathcal{G}^{m-1}| = 2^{m-1} \leq 2^d$.

Next, we use this fact to inductively augment the set $\mathcal{G}^{m-1}$ to arrive at the set $\mathcal{G}^m$ while defining the sets $A_1, \ldots, A_m$ in the process. Enumerate the classifiers in $\mathcal{G}^{m-1} = \{g_1, \ldots, g_{2^{m-1}}\}$, and let $\gamma \in \left(0, \min\left\{\varepsilon/2, 2^{1-m}\min_{i\leq m-1}\mathcal{P}(A_i^{m-1})\right\}\right)$. Suppose that, for some $k \in \{1, \ldots, 2^{m-1}\}$, there exist classifiers $g_i' \in \mathrm{B}_{\mathcal{H}^{d+1-m}}(g_i, \gamma)$ for each $i \in \{1, \ldots, k-1\}$, such that, letting $C_{k-1} = \bigcap_{i=1}^{k-1}\{x : g_i'(x) \neq g_i(x)\}$ (or $C_{k-1} = \mathcal{X}$ if $k = 1$), we have $\mathcal{P}\left(C_{k-1} \cap \bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) > 0$. Since $\mathcal{P}\left(\partial_{\tilde{\mathcal{H}}^{d+1-m}}g_k\right) > 0$, Lemma 7.25 implies there exists a sequence $\{h_t\}_{t=1}^{\infty}$ in $\tilde{\mathcal{H}}^{d+1-m}$ with $\mathcal{P}(x : h_t(x) \neq g_k(x)) \to 0$ and $\mathcal{P}\left((\limsup_{t\to\infty}\{x : h_t(x) \neq g_k(x)\}) \oplus \partial_{\tilde{\mathcal{H}}^{d+1-m}}g_k\right) = 0$; this implies the set $C = (\limsup_{t\to\infty}\{x : h_t(x) \neq g_k(x)\}) \cap C_{k-1} \cap \bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g$ satisfies $\mathcal{P}(C) = \mathcal{P}\left(C_{k-1} \cap \bigcap_{g\in\mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}}g\right) > 0$; by Lemma 7.26, $\exists t \in \mathbb{N}$ such that $\mathcal{P}(x : h_t(x) \neq g_k(x)) < \gamma$ and $\mathcal{P}(\{x : h_t(x) \neq g_k(x)\} \cap C) > 0$. Thus, defining $g_k' = h_t$ for this t, and

$C_k = \bigcap_{i=1}^{k}\{x : g_i'(x) \neq g_i(x)\}$, we have $g_k' \in \mathrm{B}_{\mathcal{H}^{d+1-m}}(g_k, \gamma)$ and $\mathcal{P}\left(C_k \cap \bigcap_{g \in \mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > 0$, thus completing the inductive step. Since we have defined $C_0 = \mathcal{X}$, and have therefore proven above that $\mathcal{P}\left(C_0 \cap \bigcap_{g \in \mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > 0$, we have established by induction that there exist classifiers $g_i' \in \mathrm{B}_{\mathcal{H}^{d+1-m}}(g_i, \gamma)$, for each $i \in \{1, \ldots, 2^{m-1}\}$, such that the set $C_{2^{m-1}} = \bigcap_{i=1}^{2^{m-1}}\{x : g_i'(x) \neq g_i(x)\}$ satisfies $\mathcal{P}\left(C_{2^{m-1}} \cap \bigcap_{g \in \mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > 0$.

Now define the set $\mathcal{G}^m = \mathcal{G}^{m-1} \cup \{g_k' : k \in \{1, \ldots, 2^{m-1}\}\}$, and for each $i \in \{1, \ldots, m-1\}$, define the set $A_i = A_i^{m-1} \cap \bigcap_{k=1}^{2^{m-1}}\{x : g_k'(x) = g_k(x)\}$; also define $A_m = \bigcap_{k=1}^{2^{m-1}}\{x : g_k'(x) \neq g_k(x)\} = C_{2^{m-1}}$. Above, we proved that $\mathcal{P}(A_m) \geq \mathcal{P}\left(C_{2^{m-1}} \cap \bigcap_{g \in \mathcal{G}^{m-1}} \partial_{\tilde{\mathcal{H}}^{d+1-m}} g\right) > 0$. Furthermore, since every $k \in \{1, \ldots, 2^{m-1}\}$ has $g_k' \in \mathrm{B}_{\mathcal{H}^{d+1-m}}(g_k, \gamma)$, by monotonicity and a union bound, every $i \in \{1, \ldots, m-1\}$ has $\mathcal{P}(A_i) = \mathcal{P}(A_i^{m-1}) - \mathcal{P}\left(A_i^{m-1} \cap \bigcup_{k=1}^{2^{m-1}}\{x : g_k'(x) \neq g_k(x)\}\right) \geq \mathcal{P}\left(A_i^{m-1}\right) - \sum_{k=1}^{2^{m-1}} \mathcal{P}\left(x : g_k'(x) \neq g_k(x)\right) \geq \mathcal{P}\left(A_i^{m-1}\right) - 2^{m-1}\gamma$; since $\gamma < 2^{1-m}\mathcal{P}(A_i^{m-1})$, we conclude $\mathcal{P}(A_i) > 0$.

Next, fix any $(x_1, \ldots, x_m) \in A_1 \times \cdots \times A_m$, and $y_1, \ldots, y_m \in \mathcal{Y}$. Since each $i \in \{1, \ldots, m-1\}$ has $A_i \subseteq A_i^{m-1}$, $(x_1, \ldots, x_{m-1}) \in A_1^{m-1} \times \cdots \times A_{m-1}^{m-1}$ as well, so that the inductive hypothesis implies $\mathcal{G}^{m-1}$ shatters $(x_1, \ldots, x_{m-1})$; therefore, some $k \in \{1, \ldots, 2^{m-1}\}$ has $\forall i \leq m-1, g_k(x_i) = y_i$. Furthermore, for this value of k, for each $i \in \{1, \ldots, m-1\}$, $A_i \subseteq \{x : g_k'(x) = g_k(x)\}$, so that $g_k'(x_i) = g_k(x_i) = y_i$ as well. On the other hand, since $A_m \subseteq \{x : g_k'(x) \neq g_k(x)\}$, we have $g_k'(x_m) \neq g_k(x_m)$; since $\mathcal{Y}$ has only two elements, we must have either $g_k'(x_m) = y_m$ or $g_k(x_m) = y_m$; in either case, $\exists g \in \mathcal{G}^m$ with $\forall i \leq m, g(x_i) = y_i$. Since $y_1, \ldots, y_m$ were arbitrary values in $\mathcal{Y}$, we conclude that $\mathcal{G}^m$ shatters $(x_1, \ldots, x_m)$, which completes the proof of the inductive step.

By the principle of induction, we have proven the claim for $n = d+1$. In particular, this implies that $\mathbb{C}$ shatters some sequence of $d+1$ points: a contradiction to the fact that $\mathrm{vc}(\mathbb{C}) = d$ (by definition). $\square$

8

A Survey of Other Topics and Techniques

The previous sections have focused on characterizing the label complexities of disagreement-based active learning methods, expressed in terms of the disagreement coefficient and VC dimension. However, there are many other aspects and approaches to the design and analysis of active learning methods. In this section, we survey a few such alternatives from the literature, though there are certainly other interesting methods not included here. Some of the topics covered focus on aspects of disagreement-based active learning not discussed in previous sections, while other topics discussed below describe alternative approaches to active learning, some of which sometimes lead to better label complexity guarantees than are possible for disagreement-based methods.

In addition to these alternative approaches to the design and analysis of active learning methods, we also discuss a few other topics from the literature on active learning. Section 8.6 describes results on the fundamental advantages of active learning over passive learning, while Section 8.7 discusses the important issue of *verifiability* of the label complexity improvements in the context of any given learning problem; this latter issue turns out to be a subtle and interesting problem for active learning, in contrast to passive learning where it is often a straightforward matter. We conclude with a brief discussion of some of the known results for active learning with classes of infinite VC dimension in Section 8.8.

For brevity, many of the results below are only accompanied by brief sketches or high-level descriptions of the respective proofs; the interested reader should refer to the original sources for detailed proofs.

8.1 Active Learning without a Version Space

The RobustCAL algorithm studied in Chapter 5 maintains a kind of soft version space V of surviving classifiers, and the mechanism for deciding whether or not to request a label Y_m is based on this set V, as is the choice of the final classifier $\hat{h}$ to output. As discussed there, this set V can be maintained *implicitly* as a set of $O(\log(m))$ constraints on the past empirical error rates of the classifiers, so that these references to the set V can be implemented as constraint satisfaction problems and constrained optimizations. However, these constraints are intuitively somewhat redundant for achieving the stated results on the label complexity, since having a relatively small empirical error rate on the set Q of *all* queried data points so far may already guarantee a relatively small error rate under $\mathcal{P}_{XY}$, as long as we have somehow managed to request the labels of every past data point in the region of disagreement of the classifiers with this low of an error rate. This last issue is somewhat tricky, and for this reason it is not presently known whether it suffices to simply *remove* the constraints (e.g., only keeping one constraint corresponding to $j = i - 1$), leaving the rest of the algorithm as is. There have been two solutions proposed to compensate for removing these constraints. One approach, due to Beygelzimer, Hsu, Langford, and Zhang [2010], building on the work of Beygelzimer, Dasgupta, and Langford [2009], uses a randomized query mechanism and includes importance weights in the calculation of empirical error rates, to compensate for the bias in the sample Q, so that it is possible to obtain rough estimates of the excess empirical error rates, even of classifiers that would not otherwise be in the soft version space, at least to the extent that we can determine they are suboptimal. A different approach, due to Hsu [2010], suggests a modification that has the additional step of also adding into the set Q the data points whose labels are *not* requested, instead providing an *inferred* label that biases toward the would-be soft version space. With either of these modifications, the only constraint needed in the constrained optimization problems that determine whether or not to request a label is a constraint on the label of the data point in question.

For brevity, we only provide the details of the latter approach described above. Specifically, the following algorithm was proposed by Hsu [2010], referred to as OracularCAL following Hsu [2010].

Algorithm: $\mathbf{OracularCAL}_\delta(n)$
0. $m \leftarrow 0$, $t \leftarrow 0$, $Q \leftarrow \{\}$
1. While $t < n$ and $m < 2^n$
2. $\quad m \leftarrow m + 1$
3. $\quad$ Let $\hat{h}_m = \mathrm{argmin}_{h \in \mathbb{C}} \mathrm{er}_Q(h)$
$\quad$ and $\hat{h}'_m = \mathrm{argmin}_{h \in \mathbb{C}: h(X_m) \neq \hat{h}_m(X_m)} \mathrm{er}_Q(h)$
4. $\quad$ If $\hat{h}'_m$ exists and $\mathrm{er}_Q(\hat{h}'_m) - \mathrm{er}_Q(\hat{h}_m) \leq U'(m-1, \delta/(m+1)^2)$
5. $\quad\quad$ Request the label Y_m, let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$, let $t \leftarrow t + 1$
6. $\quad$ Else let $Q \leftarrow Q \cup \{(X_m, \hat{h}_m(X_m))\}$
7. Return $\hat{h}_m$

Hsu [2010] shows that for any $\delta \in (0,1)$, for a particular choice of the quantity $U'(m-1, \delta/(m+1)^2)$ referenced in Step 4, this algorithm achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$, $\forall \varepsilon \in (0,1)$, letting $\varepsilon_0 = \nu + \varepsilon$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim$$
$$\left(\theta(\varepsilon_0)\frac{\nu^2}{\varepsilon^2} + \theta(\varepsilon_0)^{3/2}\frac{\nu}{\varepsilon} + \theta(\varepsilon_0)^2\right)(d + \mathrm{Log}(1/\delta))\,\mathrm{polylog}\left(\frac{d\mathrm{Log}(1/\delta)}{\varepsilon}\right).$$

Note that this is sometimes slightly worse than (5.9) from Theorem 5.4. However, it is not clear whether this bound reflects a tight analysis of OracularCAL, nor whether a refined specification of U' might improve the label complexity.

For the related method of Beygelzimer, Hsu, Langford, and Zhang [2010], the published label complexity bound, expressed in terms of ν, is slightly worse than that stated above for OracularCAL. However, they also prove a bound in terms of the parameters a and α of Condition 2.3: namely,

$$C_\alpha \theta(a\varepsilon^\alpha) a^{2-\alpha/2} \left(\frac{1}{\varepsilon}\right)^{\frac{(2-\alpha)^2}{2}} (d\mathrm{Log}(1/\varepsilon) + \mathrm{Log}(1/\delta))\mathrm{Log}\left(\frac{ad\mathrm{Log}(1/\delta)}{\varepsilon}\right)^{\frac{\alpha}{2}},$$

where C_α is an α-dependent constant. This is slightly worse than (5.8) from Theorem 5.4 (larger by roughly a factor of $1/(a\varepsilon^\alpha)^{\alpha/2}$), though often still an improvement over the bound for passive learning reflected in Theorem 3.4. A similar analysis should be possible for OracularCAL, though such an analysis has not been published at this time. Furthermore, it may even be possible to recover the bound (5.8) of Theorem 5.4 by refining the threshold U' in Step 4, though producing a formal proof establishing this remains an open problem.

Finally, we note that the reasoning that leads to OracularCAL could conceivably compose with the reasoning of Chapter 6, so that it may be possible to modify OracularCAL to use a surrogate loss ℓ, analogous to how we modified RobustCAL$_\delta$ to arrive at RobustCAL$_\delta^\ell$ in Chapter 6. Such a modification would make for a particularly elegant and computationally efficient learning algorithm, which could potentially be of substantial practical value in many applications.

8.2 The Splitting Index

As mentioned, disagreement-based active learning is certainly not the only known approach to the design and analysis of active learning. In particular, one feature lacking in disagreement-based methods is any notion of trade-off between the achieved label complexity and the number of unlabeled samples required by the algorithm. One appealing alternative to the disagreement-based approach, which *does* reflect such a trade-off, was proposed by Dasgupta [2005]. Specifically, Dasgupta [2005] proposes a type of active learning algorithm that explicitly aims at reducing the diameter of the version space by eliminating at least one classifier from each remaining pair of classifiers separated by at least a given distance. The informativeness of a given data point is then characterized by the reduction in the number of such pairs that would result from gaining knowledge of the value of $f^\star$ at that location. Given this perspective, it is natural to request the labels of the more-informative data points, and the label complexity would then be governed by just how informative these data points are. The trade-off between label complexity and number of unlabeled samples comes from the fact that these more-informative points may sometimes be quite rare compared to the less-informative points, so that we may need to examine a large number of unlabeled samples before finding such an informative point. Dasgupta [2005] characterizes this trade-off in terms of a quantity (rather, a function) referred to as the *splitting index*. Interestingly, it turns out the splitting index provides a fairly *tight* characterization of this trade-off, in a *minimax* sense, in the realizable case. The basic approach of Dasgupta [2005] can also be extended to the bounded noise setting defined in (2.2). This subsection describes the formal details of this approach to active learning, the definition of the splitting index and corresponding label complexity bound, and the extension to bounded noise.

8.2.1 The Splitting Algorithm

To describe this technique, we first need a few definitions. For any finite set $Q \subseteq \{\{h, g\} : h, g \in \mathbb{C}\}$ of unordered pairs of classifiers from $\mathbb{C}$, and any $x \in \mathcal{X}$, let $Q_x^y = \{\{h, g\} \in Q : h(x) = g(x) = y\}$ for each $y \in \mathcal{Y}$, and define

$$\mathrm{Split}(Q, x) = |Q| - \max_{y \in \mathcal{Y}} |Q_x^y|.$$

In the algorithm below, the set Q represents the pairs of surviving classifiers separated by some distance $\Delta/2$. Thus, if we hope to reduce the diameter of the version space to below $\Delta/2$, we need to eliminate at least one classifier from each of these pairs. The quantity $\mathrm{Split}(Q, x)$ represents a kind of *score* of how informative a point x is at eliminating pairs from Q, for its worst-case label. We are therefore interested in requesting the labels of points X_m with a high value of $\mathrm{Split}(Q, X_m)$. This technique is made explicit in the following algorithm due to Dasgupta [2005]. The version stated here is slightly modified compared to the original method of Dasgupta [2005], in order to express it in our present framework (i.e., active learning algorithms that take a budget n as an argument).

Algorithm: $\textbf{Splitting}_{m,\delta}(n)$
0. Let V be a minimal $\frac{\delta}{2nm}$-cover of $\mathbb{C}$; let $t = 0$,
   $\Delta = \max\{\mathcal{P}(x : h(x) \neq g(x)) : h, g \in V\}$
1. Let $Q = \{\{h, g\} \subseteq V : \mathcal{P}(x : h(x) \neq g(x)) > \Delta/2\}$
2. For $t = 1, 2, \ldots, n$
3. Let $i_t = \mathrm{argmax}_{i \in \{(t-1)m+1,\ldots,tm\}} \mathrm{Split}(Q, X_i)$
4. Request the label Y_{i_t}
5. $V \leftarrow \{h \in V : h(X_{i_t}) = Y_{i_t}\}$
6. $Q \leftarrow \{\{h, g\} \in Q : h, g \in V\}$
7. If $Q = \{\}$
8. $\Delta \leftarrow \max\{\mathcal{P}(x : h(x) \neq g(x)) : h, g \in V\}$
9. $Q \leftarrow \{\{h, g\} \subseteq V : \mathcal{P}(x : h(x) \neq g(x)) > \Delta/2\}$
10. Return any $\hat{h} \in V$ (if $|V| = 0$, return an arbitrary classifier $\hat{h}$)

Technically, the Splitting algorithm requires direct access to the distribution $\mathcal{P}$ to run. However, this is a weak kind of dependence, and can be replaced by access to a large number of unlabeled samples. Specifically, the algorithm relies on the ability to determine the distance $\mathcal{P}(x : h(x) \neq g(x))$ between any given pair of classifiers $h, g \in \mathbb{C}$; these pairwise distances can all be (uniformly) estimated up to an arbitrary precision ε (with probability $1 - \delta$) based on $O\left(\varepsilon^{-2}(d + \mathrm{Log}(1/\delta))\right)$ random unlabeled samples [Vapnik and Chervonenkis, 1971]. Furthermore, it is possible to (with probability $1 - \delta$) construct an ε-cover of $\mathbb{C}$ having near-minimal size (for use in Step 0) by using a number of unlabeled samples $O\left(\varepsilon^{-1}(d \log(1/\varepsilon) + \log(1/\delta))\right)$, simply by identifying one classifier from $\mathbb{C}$ for each of the classifications of this sample realized by classifiers in $\mathbb{C}$.

8.2.2 The Label Complexity of the Splitting Algorithm

Analysis of the label complexity of the Splitting algorithm essentially concerns the number of labels the algorithm requests before the next time the condition in Step 7 is satisfied, which indicates that the diameter of V has been (at least) halved. This number of label requests in turn depends on the value of $\mathrm{Split}(Q, X_{i_t})$ for the points X_{i_t} selected in Step 3. To quantify the guaranteed value of this, Dasgupta [2005] introduces the following definition. For any $\rho, \Delta, \tau \in (0,1)$, we say a set $\mathcal{H} \subseteq \mathbb{C}$ is (ρ, Δ, τ)-*splittable* if, for all finite $Q \subseteq \{\{h, g\} \subseteq \mathcal{H} : \mathcal{P}(x : h(x) \neq g(x)) > \Delta\}$,

$$\mathcal{P}\left(x : \mathrm{Split}(Q, x) \geq \rho|Q|\right) \geq \tau.$$

A point $x \in \mathcal{X}$ with $\mathrm{Split}(Q, x) \geq \rho|Q|$ is said to ρ-*split* Q. Recalling the definition of Split, we see that a point x that ρ-splits Q is guaranteed to eliminate at least one classifier from each of at least a ρ-fraction of the pairs in Q. Thus, the property of $\mathcal{H}$ being (ρ, Δ, τ)-splittable for relatively large values of ρ and τ would indicate that highly-informative data points are not too rare; in particular, to find a data point X_i guaranteed to eliminate a ρ-fraction of the pairs in Q, it should suffice to examine roughly $1/\tau$ of the samples. For this reason, the label complexity will be largely based on the quantity ρ, while the number of unlabeled samples used to achieve that label complexity will be related to the quantity τ. Finally, for any classifier $h \in \mathbb{C}$ and values $\varepsilon, \tau \in (0,1)$, define the *splitting index*

$$\rho_{h,\tau}(\varepsilon) = \sup\left\{\rho \in (0,1) : \forall \Delta > \varepsilon/4, \mathrm{B}(h, 4\Delta) \text{ is } (\rho, \Delta, \tau)\text{-splittable}\right\}.$$

When $h = f^\star$, abbreviate $\rho_\tau(\varepsilon) = \rho_{f^\star,\tau}(\varepsilon)$.

Dasgupta [2005] shows that, for any $\tau \leq \varepsilon/8$, every $h \in \mathbb{C}$ has $\rho_{h,\tau}(\varepsilon) \geq \varepsilon/8$. However, he also provides several examples for which $\rho_{h,\tau}(\varepsilon)$ is much larger. For threshold classifiers (Example 1), every $h \in \mathbb{C}$ has $\rho_{h,\varepsilon/4}(\varepsilon) \geq 1/2$. The reasoning is the following. For simplicity, suppose $\mathcal{P}$ is uniform in $[0,1)$ (though this easily generalizes). If $q \in \mathbb{N}$ and $Q = \{\{\mathbb{1}^{\pm}_{[z_i,1]}, \mathbb{1}^{\pm}_{[z_i',1]}\} : i \in \{1, \ldots, q\}\}$ is a set of pairs of threshold classifiers, with $\mathcal{P}(x : \mathbb{1}^{\pm}_{[z_i,1]}(x) \neq \mathbb{1}^{\pm}_{[z_i',1]}(x)) > \varepsilon/4$ for each i, and we suppose (without loss of generality) $z_i \leq z_i'$ for each i, and that $z_i \leq z_{i+1}$ for each $i \in \{1, \ldots, q-1\}$, then any point $x \in [z_{\lceil q/2 \rceil}, z_{\lceil q/2 \rceil} + \varepsilon/4)$ will eliminate at least $\lceil q/2 \rceil$ pairs from Q regardless of its label: contradicting $\mathbb{1}^{\pm}_{[z_1,1]}, \ldots, \mathbb{1}^{\pm}_{[z_{\lceil q/2 \rceil},1]}$ if labeled -1, or else contradicting $\mathbb{1}^{\pm}_{[z'_{\lceil q/2 \rceil},1]}, \ldots, \mathbb{1}^{\pm}_{[z_q',1]}$ if labeled $+1$ (noting that each $i \geq \lceil q/2 \rceil$ has $z_i' \geq z_i + \varepsilon/4 \geq z_{\lceil q/2 \rceil} + \varepsilon/4$). Dasgupta [2005] generalizes this to the class of homogeneous linear separators in $\mathbb{R}^k$, for $k \in \mathbb{N}$, under $\mathcal{P}$ a uniform distribution on the surface of an origin-centered sphere, showing that for a value of $\tau \propto \varepsilon$, every homogeneous linear separator h has $\rho_{h,\tau}(\varepsilon) \geq 1/4$.

Based on the above definition of the splitting index, Dasgupta [2005] proves a variant of the following theorem, bounding the label complexity of the Splitting algorithm.

Theorem 8.1. For any $\varepsilon, \delta \in (0,1)$ and $\tau \in (0, \varepsilon/2)$, for $m = \lceil 1/\tau \rceil$, $\text{Splitting}_{m,\delta}$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ in the realizable case,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \frac{d}{\rho_\tau(\varepsilon)} \mathrm{Log}\left(\frac{d}{\delta\varepsilon\tau\rho_\tau(\varepsilon)}\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Proof Sketch. For brevity, we only provide a brief sketch of the proof here. Let V_0 denote the initial value of the set V in the algorithm. Note that the choice of V_0 as a $\delta/(2nm)$-cover of $\mathbb{C}$ guarantees that each of the mn unlabeled samples used by the algorithm has only a $\delta/(2nm)$ probability of falling in the region where $f^\star$ disagrees with its closest representative h_0 in V_0, so that a union bound implies that $h_0 \in V$ in the end with probability at least $1 - \delta/2$. The essential idea of the rest of the proof is that we maintain the invariant that $V \subseteq \mathrm{B}(h_0, \Delta)$, so that on any given iteration of the algorithm with $\Delta > \varepsilon/2$, we have $V \subseteq \mathrm{B}(f^\star, \Delta + \varepsilon/2) \subseteq \mathrm{B}(f^\star, 2\Delta)$, and therefore the definition of $\rho_\tau(\varepsilon)$ implies that, with probability at least $1 - (1-\tau)^m \geq 1 - e^{-\tau m} \geq 3/5$, we should encounter a $\rho_\tau(\varepsilon)$-splitting point for Q among the next m samples. By a Chernoff bound (after some additional reasoning about dependences), with probability at least $1 - \delta/2$, if there are

$$\frac{40}{3}\mathrm{Log}\,(2/\delta) \vee \frac{10}{3\rho_\tau(\varepsilon)}\mathrm{Log}\left(|V_0|^2\right) \lceil \log_2(2/\varepsilon) \rceil \tag{8.1}$$

iterations of the loop in the algorithm prior to halting or reaching a $\Delta \leq \varepsilon/2$, then the algorithm will have at least $\frac{1}{\rho_\tau(\varepsilon)}\mathrm{Log}\left(|V_0|^2\right) \lceil \log_2(2/\varepsilon) \rceil$ times t in which $\mathrm{Split}(Q, X_{i_t}) \geq \rho_\tau(\varepsilon)|Q|$. Since each such $\rho_\tau(\varepsilon)$-splitting point is guaranteed to eliminate at least a $\rho_\tau(\varepsilon)$-fraction of the pairs in Q, the condition in Step 7 will be satisfied at least once every $\frac{1}{\rho_\tau(\varepsilon)}\mathrm{Log}\left(|V_0|^2\right)$ times we query a $\rho_\tau(\varepsilon)$-splitting point. Since

Δ is at least halved each time this happens, we need only satisfy this condition at most $\lceil \log_2(2/\varepsilon) \rceil$ times before $\Delta \leq \varepsilon/2$; therefore, the above number of queries of $\rho_\tau(\varepsilon)$-splitting points suffices to guarantee $\Delta \leq \varepsilon/2$. In particular, upon achieving $\Delta \leq \varepsilon/2$, we have $V \subseteq \mathrm{B}(h_0, \Delta) \subseteq \mathrm{B}(f^\star, \Delta + \varepsilon/2) \subseteq \mathrm{B}(f^\star, \varepsilon)$, so that we are guaranteed $\mathrm{er}(\hat{h}) \leq \varepsilon$. Since $|V_0| \lesssim d\left(32e\left(\frac{nm}{\delta} \vee \frac{1}{\varepsilon}\right)\right)^d$ [Haussler, 1995], the above number of iterations (8.1) sufficient to reach this stage is

$$\lesssim \frac{d}{\rho_\tau(\varepsilon)} \mathrm{Log}\left(\frac{n}{\delta\varepsilon\tau}\right) \mathrm{Log}\left(1/\varepsilon\right).$$

Thus, taking any n greater than this suffices to guarantee the final V set is contained in $\mathrm{B}(f^\star, \varepsilon)$. In particular, taking n at least as large as the bound given in the theorem statement (with appropriate constant factors) suffices to satisfy this. A union bound to combine the above two $(1 - \delta/2)$-probability events completes the proof. $\square$

Also note that, since the algorithm processes exactly m unlabeled points for every one label it requests, we see that with probability at least $1 - \delta$, the number of unlabeled samples needed by the Splitting algorithm to achieve error rate ε is at most

$$m\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \frac{d}{\tau\rho_\tau(\varepsilon)} \mathrm{Log}\left(\frac{d}{\delta\varepsilon\tau\rho_\tau(\varepsilon)}\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Thus, $\rho_\tau(\varepsilon)$ and τ describe a trade-off between the label complexity and the number of unlabeled samples used by the algorithm, a feature that was not present in disagreement-based active learning.

8.2.3 The Minimax Label Complexity in the Realizable Case

It turns out the splitting index can also be used to prove lower bounds on the worst-case label complexity of *any* active learning algorithm with a bounded number of unlabeled samples, in the realizable case: that is, lower bounds on a certain *minimax* label complexity. In particular, this implies that (aside from constants and logarithmic factors), the worst-case label complexity of the Splitting algorithm (in the realizable case) is typically nearly optimal, among this family of algorithms. To make this formal, for any $\varepsilon, \delta \in (0,1)$, let us denote by

$$\Lambda^*(\varepsilon, \delta; \mathbb{C}, \mathcal{P}) = \inf_{\Lambda} \sup_{P} \Lambda(\varepsilon, \delta, P),$$

where P ranges over all realizable-case distributions having $\mathcal{P}$ as their marginal distribution over $\mathcal{X}$, and Λ ranges over all label complexity functions achieved by active learning algorithms $\mathcal{A}$ with the property that, for every $n \in \mathbb{N}$, every label Y_t requested during the execution of $\mathcal{A}(n)$ is guaranteed to have $t \leq M_n$, for some $(n, \mathcal{A})$-dependent constant $M_n \in \mathbb{N}$ (independent of $\mathcal{P}_{XY}$); note that, since $\mathcal{P}$ is fixed here, and P is restricted to the realizable case, we can equivalently think of $\sup_P \Lambda(\varepsilon, \delta, P)$ as being the value of $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$ for $\mathcal{P}_{XY}$ with a worst-case choice of *target function* $f^\star \in \mathbb{C}$; also note that M_n may be quite large (e.g., it is 2^n in the CAL algorithm stated in Chapter 5), so that this restriction on Λ should be considered quite mild (though it might certainly be very interesting to understand how the minimax label complexity changes when this restriction is removed).

The following result is based on a lower-bound argument of Dasgupta [2005] [see also Balcan and Hanneke, 2012], combined with the limiting case of Theorem 8.1.

Theorem 8.2. For any $\varepsilon \in (0, 1/4)$, $\delta \in (0, 3/16)$, and marginal distribution $\mathcal{P}$ over $\mathcal{X}$,

$$\inf_{\tau > 0} \sup_{h \in \mathbb{C}} \frac{1}{\rho_{h,\tau}(8\varepsilon)} \lesssim \Lambda^*(\varepsilon, \delta; \mathbb{C}, \mathcal{P}) \lesssim \inf_{\tau > 0} \sup_{h \in \mathbb{C}} \frac{d}{\rho_{h,\tau}(\varepsilon)} \mathrm{Log}\left(\frac{d}{\varepsilon \delta \tau \rho_{h,\tau}(\varepsilon)}\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Proof. The upper bound follows as a limiting case of the bound in Theorem 8.1, since the Splitting algorithm uses at most mn unlabeled samples, regardless of the distribution. To prove the lower bound, suppose Λ is a label complexity achieved by some active learning algorithm $\mathcal{A}$ satisfying the requirement from the definition of Λ^* (regarding the existence of the $\{M_n\}_{n \in \mathbb{N}}$ sequence); for the sake of contradiction, suppose $\sup_P \Lambda(\varepsilon, \delta, P) < \left\lfloor \inf_{\tau > 0} \sup_{h \in \mathbb{C}} \frac{1}{8\rho_{h,\tau}(8\varepsilon)} \right\rfloor$, where P ranges over distributions over $\mathcal{X} \times \mathcal{Y}$ in the realizable case having marginal distribution $\mathcal{P}$ over $\mathcal{X}$. Denote $n = \left\lfloor \inf_{\tau > 0} \sup_{h \in \mathbb{C}} \frac{1}{8\rho_{h,\tau}(8\varepsilon)} \right\rfloor$.

Let $\tau \in (0, \varepsilon)$ be such that $M_n < 1/(4\tau)$, and let $f \in \mathbb{C}$ be such that $\rho_{f,\tau}(8\varepsilon) < 2\inf_{h\in\mathbb{C}} \rho_{h,\tau}(8\varepsilon)$; such a classifier f is guaranteed to exist since (as mentioned above) any $\tau \in (0, \varepsilon)$ has $\inf_{h\in\mathbb{C}} \rho_{h,\tau}(8\varepsilon) \geq \varepsilon > 0$. In particular, note that $n < \frac{1}{4\rho_{f,\tau}(8\varepsilon)}$. For brevity, denote $\rho = \rho_{f,\tau}(8\varepsilon)$. Let $\Delta > 2\varepsilon$ and finite $Q \subseteq \{\{h, g\} \subseteq \mathrm{B}(f, 4\Delta) : \mathcal{P}(x : h(x) \neq g(x)) > \Delta\}$ be such that $\mathcal{P}(x : \mathrm{Split}(Q, x) \geq 2\rho|Q|) < \tau$.

For each $x \in \mathcal{X}$, let $g^*(x) = \mathrm{argmax}_{y\in\mathcal{Y}} |Q_x^y|$. For any classifier g, let $\hat{h}_g$ and $\mathcal{L}_g$ denote the classifier returned by $\mathcal{A}(n)$ and the set of indices of the labels that would be requested during execution of $\mathcal{A}(n)$, respectively, if $X_1, X_2, \ldots$ are defined as usual (i.i.d. $\mathcal{P}$), but $Y_m = g(X_m)$ instead of $f^\star(X_m)$, for all $m \in \mathbb{N}$. Thus, $\hat{h}_g$ and $\mathcal{L}_g$ are random variables depending only on the $X_1, X_2, \ldots$ sequence (and any independent internal randomness in the algorithm $\mathcal{A}$); furthermore, aside from independent internal randomness, the behavior of $\mathcal{A}$ is completely determined by the sequence $X_1, X_2, \ldots$ and the sequence of values Y_m requested, and therefore $\hat{h}_g = \hat{h}_{g'}$ for any classifiers g and g' with $\forall m \in \mathcal{L}_g, g'(X_m) = g(X_m)$.

We therefore have

$$\begin{aligned}
&\sup_{f\in\mathbb{C}} \mathbb{P}(\mathcal{P}(x : \hat{h}_f(x) \neq f(x)) > \varepsilon) \\
&\geq \frac{1}{2|Q|} \sum_{\{h_1,h_2\}\in Q} \sum_{i\in\{1,2\}} \mathbb{P}(\mathcal{P}(x : \hat{h}_{h_i}(x) \neq h_i(x)) > \varepsilon) \\
&= \mathbb{E}\left[\frac{1}{2|Q|} \sum_{\{h_1,h_2\}\in Q} \sum_{i\in\{1,2\}} \mathbb{1}\left[\mathcal{P}(x : \hat{h}_{h_i}(x) \neq h_i(x)) > \varepsilon\right]\right] \\
&\geq \mathbb{E}\left[\frac{1}{2|Q|} \sum_{\{h_1,h_2\}\in Q} \sum_{i\in\{1,2\}} \mathbb{1}\left[\mathcal{P}(x : \hat{h}_{g^*}(x) \neq h_i(x)) > \varepsilon\right]\right. \\
&\qquad\qquad\left. \times \prod_{m\in\mathcal{L}_{g^*}} \mathbb{1}[h_i(X_m) = g^*(X_m)]\right]. \qquad (8.2)
\end{aligned}$$

Note that, for any $x \in \mathcal{X}$ with $\mathrm{Split}(Q, x) < 2\rho|Q|$, there are greater than $(1 - 2\rho)|Q|$ pairs $\{h_1, h_2\} \in Q$ for which $h_1(x) = h_2(x) = g^*(x)$. Therefore, if every $m \in \mathcal{L}_{g^*}$ has $\mathrm{Split}(Q, X_m) < 2\rho|Q|$, there are at least $(1-2\rho n)|Q| \geq |Q|/2$ pairs $\{h_1, h_2\} \in Q$ for which $\forall m \in \mathcal{L}_{g^*}, h_1(X_m) =$

$h_2(X_m) = g^*(X_m)$. Recalling that $\mathcal{L}_{g^*} \subseteq \{1, \ldots, M_n\}$, this will be satisfied whenever every $m \leq M_n$ has $\mathrm{Split}(Q, X_m) < 2\rho|Q|$. Furthermore, since $\Delta > 2\varepsilon$, every classifier g has either $\mathcal{P}(x : g(x) \neq h_1(x)) > \varepsilon$ or $\mathcal{P}(x : g(x) \neq h_2(x)) > \varepsilon$ for each $\{h_1, h_2\} \in Q$, so that $\sum_{i \in \{1,2\}} \mathbb{1}[\mathcal{P}(x : \hat{h}_{g^*}(x) \neq h_i(x)) > \varepsilon] \geq 1$. We therefore have that (8.2) is at least as large as

$$\mathbb{E}\left[\frac{1}{2|Q|}\frac{|Q|}{2}\prod_{m=1}^{M_n}\mathbb{1}\left[\mathrm{Split}(Q, X_m) < 2\rho|Q|\right]\right] =$$
$$\frac{1}{4}\mathcal{P}(x : \mathrm{Split}(Q, x) < 2\rho|Q|)^{M_n} > \frac{1}{4}(1-\tau)^{M_n} > \frac{1}{4}(1-\tau)^{1/(4\tau)} \geq \frac{3}{16},$$

where the first equality is due to independence of the X_m variables. We have thus reached a contradiction, since our choice of n (and the definition of Λ) imply that $\sup_{g \in \mathbb{C}} \mathbb{P}(\mathcal{P}(x : \hat{h}_g(x) \neq g(x)) > \varepsilon) \leq \delta \leq 3/16$. $\square$

Reflecting on Theorem 8.2, we see that the splitting index provides a fairly good characterization of the minimax label complexity. However, we should note that, even ignoring the factor of d, and factors of $\mathrm{Log}(1/\varepsilon)$ and $\mathrm{Log}(1/\delta)$, there remains a factor of $\mathrm{Log}(1/\tau)$ in the upper bound that is not present in the lower bound, so that the value of $\rho_{h,\tau}(\varepsilon)$ at the value of τ minimizing the upper bound does not necessarily match that of the lower bound (which is realized in the limit as $\tau \to 0$): that is, we cannot quite say that $1/\rho_{h,\tau}(\varepsilon)$ tightly characterizes the minimax label complexity, since it may have different values in the upper bound compared to the lower bound. Resolving the issue of this extra factor of $\mathrm{Log}(1/\tau)$ remains an important open problem in the theory of active learning. That said, in many cases it happens that $\rho_{h,\tau_0}(\varepsilon)$ is within a constant factor of $\lim_{\tau \to 0} \rho_{h,\tau}(\varepsilon)$, even for some value of $\tau_0 = \mathrm{poly}(\varepsilon)$ (as it is in the examples given above), so that $\mathrm{Log}(1/\tau_0) = O(\mathrm{Log}(1/\varepsilon))$ anyway; in these cases, we can consider the $1/\rho_{h,\tau}(\varepsilon)$ values in the upper and lower bounds to be roughly the same, thus providing a fairly tight characterization of the minimax label complexity in the realizable case.

8.2.4 Extension to Bounded Noise

Theorem 8.1 applies to the realizable case only. However, it is possible to extend the splitting technique beyond the realizable case: specifically, to the *bounded noise* case. Formally, as mentioned in Section 2.3, we say a distribution $\mathcal{P}_{XY}$ satisfies a bounded noise condition if $f^{\star}(\cdot) = \mathrm{sign}(\eta(\cdot) - 1/2)$ (i.e., the Bayes optimal classifier) and (2.2) is satisfied for some constant $a \in [1, \infty)$: that is, $\mathcal{P}(x : |\eta(x) - 1/2| < 1/(2a)) = 0$. The following algorithm is a noise-robust variant of the Splitting algorithm, for which we will provide an analysis under the bounded noise condition.

Algorithm: **RobustSplitting**$_{m,\delta}(n)$
0. Let V be a minimal $\frac{\delta}{3nm}$-cover of $\mathbb{C}$;
   let $\Delta = \max\{\mathcal{P}(x : h(x) \neq g(x)) : h, g \in V\}$, $v = |V|$
1. For every $h, g \in V$, let $M_{hg} = 0$
2. Let $Q = \{\{h, g\} \subseteq V : \mathcal{P}(x : h(x) \geq g(x)) > \Delta/2\}$
3. For $t = 1, 2, \ldots, n$
4. Let $i_t = \mathrm{argmax}_{i \in \{(t-1)m+1, \ldots, tm\}} \mathrm{Split}(Q, X_i)$
5. Request the label Y_{i_t}
6. $\forall h, g \in V$ with $h(X_{i_t}) \neq g(X_{i_t}) = Y_{i_t}$, let $M_{hg} \leftarrow M_{hg} + 1$
7. $V \leftarrow \Big\{h \in V : \forall g \in V, M_{hg} - M_{gh} \leq$
   $4\sqrt{M_{hg} \ln(3vn/\delta)} + 2\log_2(3vn/\delta)\Big\}$
8. $Q \leftarrow \{\{h, g\} \in Q : h, g \in V\}$
9. If $Q = \{\}$
10. $\Delta \leftarrow \max\{\mathcal{P}(x : h(x) \neq g(x)) : h, g \in V\}$
11. $Q \leftarrow \{\{h, g\} \subseteq V : \mathcal{P}(x : h(x) \neq g(x)) > \Delta/2\}$
12. Return any $\hat{h} \in V$ (if $|V| = 0$, return an arbitrary classifier $\hat{h}$)

The following theorem provides a label complexity bound for RobustSplitting under a bounded noise condition. It follows by a combination of the reasoning given above for the Splitting algorithm and a basic application of Chernoff and union bounds to guarantee that we only remove a classifier from V if we can determine with high confidence that it is worse than some other classifier in V, based on a direct comparison of those two classifiers. A similar result appears in the recent work of Balcan and Hanneke [2012].

Theorem 8.3. For any $\varepsilon, \delta \in (0,1)$ and $\tau \in (0, \varepsilon/2)$, for $m = \lceil 1/\tau \rceil$, $\text{RobustSplitting}_{m,\delta}$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ with $\text{sign}(\eta - 1/2) \in \mathbb{C}$ that satisfies (2.2) with a given value a,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \frac{a^2 d^2}{\rho_\tau(\varepsilon)} \mathrm{Log}\left(\frac{ad}{\varepsilon\delta\tau\rho_\tau(\varepsilon)}\right)^2 \mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Proof Sketch. The proof is in two parts. First, letting V_0 denote the initial set V, and $h_0 = \mathrm{argmin}_{h \in V} \mathcal{P}(x : h(x) \neq f^\star(x))$, we show that $h_0 \in V$ is maintained as an invariant, with probability $1 - 2\delta/3$. After establishing this, we proceed to show that taking n as in the bound in the theorem statement suffices to reach a value of $\Delta \leq \varepsilon/2$.

Since V is a $\frac{\delta}{3nm}$-cover of $\mathbb{C}$, we know $\mathcal{P}(x : h_0(x) \neq f^\star(x)) \leq \frac{\delta}{3nm}$. Thus, by a union bound, with probability at least $1 - \delta/3$, every $i \in \{1, \ldots, nm\}$ satisfies $h_0(X_i) = f^\star(X_i)$; in particular, this is the case for the n values $i_1, \ldots, i_n$ corresponding to the requested labels. For any $h, g \in \mathbb{C}$ and $t \in \{1, \ldots, n\}$, denote $\hat{M}_{hg}^{(t)} = \sum_{k=1}^{t} \mathbb{1}[h(X_{i_k}) \neq Y_{i_k}] \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X_{i_k})$ and $\hat{M}_{hg}^{(0)} = 0$; also denote by $\hat{M}_h^{(t)} = \hat{M}_{hf^\star}^{(t)} + \hat{M}_{f^\star h}^{(t)} = |\{X_{i_1}, \ldots, X_{i_t}\} \cap \mathrm{DIS}(\{h, f^\star\})|$. Since $\mathbb{P}(Y = f^\star(X)|X) \geq \frac{1}{2} + \frac{1}{2a} > 1/2$ for $(X,Y) \sim \mathcal{P}_{XY}$, a Chernoff bound (along with some additional reasoning to address dependences; see Balcan and Hanneke, 2012, for a related discussion) implies that, for any $t \in \{1, \ldots, n\}$ and $h \in V_0$, with probability at least $1 - \delta/(3vn)$,

$$\hat{M}_{f^\star h}^{(t)} - \hat{M}_{hf^\star}^{(t)} \leq -\frac{1}{a}\hat{M}_h^{(t)} + \sqrt{8(1-1/a)\hat{M}_h^{(t)} \ln(3vn/\delta)} + 2\log_2(3vn/\delta). \tag{8.3}$$

By a union bound, this holds simultaneously for all $t \in \{1, \ldots, n\}$ and all $h \in V_0$ with probability at least $1 - \delta/3$. Furthermore, since $\hat{M}_{f^\star h}^{(t)} - \hat{M}_{hf^\star}^{(t)} \leq 0$ when $\hat{M}_{f^\star h}^{(t)} < \hat{M}_{hf^\star}^{(t)}$, and $\hat{M}_h^{(t)} \leq 2\max\{\hat{M}_{hf^\star}^{(t)}, \hat{M}_{f^\star h}^{(t)}\} = 2\hat{M}_{f^\star h}^{(t)}$ when $\hat{M}_{f^\star h}^{(t)} \geq \hat{M}_{hf^\star}^{(t)}$, (8.3) implies

$$\hat{M}_{f^\star h}^{(t)} - \hat{M}_{hf^\star}^{(t)} \leq 4\sqrt{\hat{M}_{f^\star h}^{(t)} \ln(3vn/\delta)} + 2\log_2(3vn/\delta).$$

In particular, combining this with the first event, which implies $\hat{M}_{f^\star h}^{(t)} = \hat{M}_{h_0 h}^{(t)}$ and $\hat{M}_{hf^\star}^{(t)} = \hat{M}_{hh_0}^{(t)}$, we see that with with probability at least

$1 - 2\delta/3$, $\forall t \in \{1, \ldots, n\}$, $\forall h \in V_0$,

$$\hat{M}_{h_0 h}^{(t)} - \hat{M}_{h h_0}^{(t)} \leq 4\sqrt{\hat{M}_{h_0 h}^{(t)} \ln(3vn/\delta)} + 2\log_2(3vn/\delta).$$

Since the value $M_{h_0 g}$ is equal to $\hat{M}_{h_0 g}^{(t)}$ as long as $h_0, g \in V$ in the algorithm, we see that with probability at least $1 - 2\delta/3$, $h_0 \in V$ is maintained as an invariant throughout the execution of the algorithm.

Next, note that (8.3) implies

$$\hat{M}_{hf^\star}^{(t)} - \hat{M}_{f^\star h}^{(t)} \geq \frac{1}{a}\hat{M}_h^{(t)} - \sqrt{8(1-1/a)\hat{M}_h^{(t)} \ln(3vn/\delta)} - 2\log_2(3vn/\delta).$$

Therefore, on the above combined event of probability $1 - 2\delta/3$, any h remaining in V after the update in Step 7 satisfies

$$\begin{aligned} &4\sqrt{\hat{M}_{hf^\star}^{(t)} \ln(3vn/\delta)} + 2\log_2(3vn/\delta) \\ &\qquad \geq \frac{1}{a}\hat{M}_h^{(t)} - \sqrt{8(1-1/a)\hat{M}_h^{(t)} \ln(3vn/\delta)} - 2\log_2(3vn/\delta). \end{aligned}$$

Since $\hat{M}_h^{(t)} \geq \hat{M}_{hf^\star}^{(t)}$, the above inequality implies

$$256a^2 \ln(3vn/\delta) > \hat{M}_h^{(t)}.$$

For each $t \in \{1, \ldots, n\}$, there are at least $\mathrm{Split}(Q, X_{i_t})$ pairs $\{h, g\} \in Q$ for which either $\hat{M}_h^{(t)} = \hat{M}_h^{(t-1)} + 1$ or $\hat{M}_g^{(t)} = \hat{M}_g^{(t-1)} + 1$. In particular, combined with the above inequality constraining the values of $\hat{M}_h^{(t)}$ for classifiers in V, this implies we require at most $\frac{512a^2}{\rho_\tau(\varepsilon)} \ln\left(v^2\right) \ln(3vn/\delta)$ values of i_t with $\mathrm{Split}(Q, X_{i_t}) \geq \rho_\tau(\varepsilon)|Q|$ between times when the condition in Step 9 is satisfied. Furthermore, as in the proof of Theorem 8.1, since $h_0 \in V$ is maintained as an invariant, and $\mathcal{P}(x : h_0(x) \neq f^\star(x)) \leq \varepsilon/2$, we also maintain the invariant that $V \subseteq \mathrm{B}(h_0, \Delta) \subseteq \mathrm{B}(f^\star, \Delta + \varepsilon/2)$, so that if the algorithm reaches a value of $\Delta \leq \varepsilon/2$, we have $V \subseteq \mathrm{B}(f^\star, \varepsilon)$, and therefore $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \mathcal{P}(x : \hat{h}(x) \neq f^\star(x)) \leq \varepsilon$. Since Δ is at least halved every time the condition in Step 7 is satisfied, we see that this will be achieved within the first $\frac{512a^2}{\rho_\tau(\varepsilon)} \ln\left(v^2\right) \ln(3vn/\delta)\lceil\log_2(2/\varepsilon)\rceil$ times t in which $\mathrm{Split}(Q, X_{i_t}) \geq \rho_\tau(\varepsilon)|Q|$. By the same reasoning as in the proof of Theorem 8.1, with probability at least $1 - \delta/3$, if there are

$$\frac{5120a^2}{3\rho_\tau(\varepsilon)} \ln\left(v^2\right) \ln(3vn/\delta)\lceil \log_2(2/\varepsilon)\rceil$$

iterations of the loop in the RobustSplitting$_{m,\delta}$ algorithm prior to halting or reaching a value of $\Delta \leq \varepsilon/2$, then we will have encountered at least this many $\rho_\tau(\varepsilon)$-splitting points, and therefore achieve $\Delta \leq \varepsilon/2$, and hence $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$. By a union bound to combine the above three events, each having probability at least $1-\delta/3$, we have that this will happen with probability at least $1 - \delta$, as long as

$$n \geq \frac{5120a^2}{3\rho_\tau(\varepsilon)} \ln\left(v^2\right) \ln(3vn/\delta)\lceil \log_2(2/\varepsilon)\rceil.$$

Since $v \lesssim d\left(32e\left(\frac{nm}{\delta} \vee \frac{1}{\varepsilon}\right)\right)^d$ [Haussler, 1995], we see that, for an appropriate choice of constant factors, this will be satisfied for any n at least the size of the bound indicated in the theorem statement. $\square$

The bound in Theorem 8.3 is essentially similar to that of Theorem 8.1, except for an increase by a factor of roughly da^2 and an additional logarithmic factor. It is possible to eliminate the extra factor of d by a slightly different algorithm [Kääriäinen, 2006], but it is not known whether this is possible without a significant increase in the required number of unlabeled data points, or more specifically, while preserving the elegant description of the *trade-off* between the label complexity and the number of unlabeled data points given by the balance between $\rho_\tau(\varepsilon)$ and τ in the splitting index analysis.

8.2.5 The Splitting Index and the Disagreement Coefficient

Since we have now seen two distinct techniques for bounding the label complexity of active learning, namely the splitting and disagreement approaches, it is natural to ask whether they are formally related at some basic level. At present, there are no published results formally relating $\rho_\tau(\varepsilon)$ and $\theta(\varepsilon)$ in the general case. At a less-formal level, we may observe that the splitting index includes a trade-off that allows it to describe the informativeness of rare points that we only expect to appear in large data sets of size $\tilde{\Omega}(1/\tau)$. The lack of such a trade-off in the disagreement coefficient suggests that $1/\rho_\tau(\varepsilon)$ may often be smaller than $\theta(\varepsilon)$ when τ is taken sufficiently small.

The following example illustrates an extreme case of this issue (i.e., rare-but-informative points). Consider the case of $\mathcal{X} = [0,3)$, $\mathbb{C} = \{\mathbb{1}^{\pm}_{[a,b)\cup[1,1+a)\cup[2,2+b)} : 0 \leq a \leq b < 1\}$, and $\mathcal{P}_{XY}$ in the realizable case, with $f^\star = \mathbb{1}^{\pm}_{\emptyset} = \mathbb{1}^{\pm}_{[0,0)\cup[1,1)\cup[2,2)}$, and with $\mathcal{P}$ defined by $\mathcal{P}(A) = (1-\varepsilon)\lambda(A\cap[0,1)) + (\varepsilon/2)\lambda(A\cap[1,3))$ for any measurable $A \subseteq \mathcal{X}$, where λ is the Lebesgue measure. In this case, for any $r \in (\varepsilon, 1)$, $\mathrm{B}(f^\star, r) = \{\mathbb{1}^{\pm}_{[a,b)\cup[1,1+a)\cup[2,2+b)} : 0 \leq a \leq b < 1, (1-\varepsilon)(b-a) + (\varepsilon/2)(a+b) \leq r\}$, so that $\mathrm{DIS}(\mathrm{B}(f^\star, r)) = [0,3)$, and hence $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r))) = 1$, which implies $\theta(\varepsilon) = 1/\varepsilon$. However, fix any finite set $Q \subseteq \{\{h,g\} \subseteq \mathbb{C} : \mathcal{P}(x : h(x) \neq g(x)) > \varepsilon/4\}$, and enumerate the elements of Q as $\{\mathbb{1}^{\pm}_{[z_{i1},z_{i2})\cup[1,1+z_{i1})\cup[2,2+z_{i2})}, \mathbb{1}^{\pm}_{[z'_{i1},z'_{i2})\cup[1,1+z'_{i1})\cup[2,2+z'_{i2})}\}_{i=1}^{q}$, where $q = |Q|$. For any $i \leq q$, since the corresponding pair of classifiers is $(\varepsilon/4)$-separated, we must have either $|z_{i1} - z'_{i1}| > \varepsilon/8$ or $|z_{i2} - z'_{i2}| > \varepsilon/8$. Therefore, if we let $m_1 = |\{i \in \{1,\ldots,q\} : |z_{i1} - z'_{i1}| \geq |z_{i2} - z'_{i2}|\}|$, $m_2 = q - m_1$, and let $\{i_{t1}\}_{t=1}^{m_1}$ be the subsequence of indices $i \in \{1,\ldots,q\}$ for which $|z_{i1} - z'_{i1}| \geq |z_{i2} - z'_{i2}|$, and $\{i_{t2}\}_{t=1}^{m_2}$ the subsequence of indices $i \in \{1,\ldots,q\}$ for which $|z_{i2} - z'_{i2}| > |z_{i1} - z'_{i1}|$, then we have $\max\{m_1, m_2\} \geq \lceil q/2 \rceil$, and for $k = \mathrm{argmax}_{j\in\{1,2\}} m_j$, $\min_{t\in\{1,\ldots,m_k\}} |z_{i_{tk}k} - z'_{i_{tk}k}| > \varepsilon/8$. Without loss of generality, we may suppose each $t \in \{1,\ldots,m_k\}$ has $z_{i_{tk}k} < z'_{i_{tk}k}$, and that $z_{i_{tk}k}$ is nondecreasing in t. Then for $T = \lceil m_k/2 \rceil$, and any $x \in [k + z_{i_{Tk}k}, k + z_{i_{Tk}k} + \varepsilon/8)$, the T classifiers corresponding to the $z_{i_{1k}k}, \ldots, z_{i_{Tk}k}$ values classify x as -1, while the $m_k - T + 1$ classifiers corresponding to the $z'_{i_{Tk}k}, \ldots, z'_{i_{m_k k}k}$ values classify x as $+1$ (since each $t \in \{T,\ldots,m_k\}$ has $z'_{i_{tk}k} > z_{i_{tk}k} + \varepsilon/8 \geq z_{i_{Tk}k} + \varepsilon/8 > x$); this implies $\mathrm{Split}(Q,x) \geq \min\{T, m_k - T + 1\} = \lceil m_k/2 \rceil \geq q/4$. Since $\mathcal{P}([k + z_{i_{Tk}k}, k + z_{i_{Tk}k} + \varepsilon/8)) = \varepsilon^2/16$, and the above reasoning holds for any set Q, we have that for any $\tau \leq \varepsilon^2/16$, $\rho_\tau(\varepsilon) \geq 1/4$. Thus, although $\theta(\varepsilon) = 1/\varepsilon$, we have $1/\rho_\tau(\varepsilon) \leq 4$.

8.3 Combinatorial Dimensions

We have now seen two distinct approaches to the design and analysis of active learning: namely, the disagreement-based approach and the splitting-based approach. In this subsection, we review another approach, rooted in the classic literature on Exact learning with membership queries. We begin by reviewing this related topic, along with the relevant known results for that setting. We then discuss an extension of these ideas to our present active learning setting.

8.3.1 Exact Learning

There is a thread in the classic learning theory literature that explores a rather extreme variant of active learning: namely, *Exact* learning with membership queries. Specifically, in this setting, there is a target function $f^\star \in \mathbb{C}$, and a learning algorithm $\mathcal{A}$ may select *any* point $x \in \mathcal{X}$ and request to observe the corresponding label $f^\star(x)$; this is the only information about $f^\star$ the algorithm has access to. An algorithm $\mathcal{A}$ of this type is called an *MQ-algorithm for Exact learning* $\mathbb{C}$ if, for any $f^\star \in \mathbb{C}$, after some number of these queries, it returns the classifier $f^\star$. This is clearly a much stronger requirement than we have been discussing above, since the algorithm is not allowed a δ failure probability, and furthermore must return *exactly* the target function, rather than merely any classifier ε-close to it. However, this stronger requirement is balanced to some extent by a stronger querying capability: namely, the ability to request the target function's label at *any* location, rather than restricting such queries to a given pool of unlabeled data. The main quantity of interest in the literature on this topic is the *minimax* query complexity, $\Lambda^*_{\mathrm{MQ}}(\mathbb{C})$, defined as the smallest integer q such that there exists an MQ-algorithm for Exact learning $\mathbb{C}$ with the guarantee that, for every $f^\star \in \mathbb{C}$, the algorithm makes at most q queries before returning $f^\star$. In particular, due to the strength of the Exact learning requirement, finite values of $\Lambda^*_{\mathrm{MQ}}(\mathbb{C})$ are only achievable for *finite* hypothesis classes $\mathbb{C}$.

Within this setting, Hegedüs [1995] defines a complexity measure $\mathrm{XTD}(\mathbb{C})$, called the *extended teaching dimension* (due to its relation to an earlier quantity, called the *teaching dimension*, used by Goldman and Kearns, 1995, to study the complexity of a kind of exact teaching). Specifically, for any $h : \mathcal{X} \to \mathcal{Y}$, $m \in \mathbb{N}$, and $S = (x_1, \ldots, x_m) \in \mathcal{X}^m$, let $h(S) = (h(x_1), \ldots, h(x_m))$. Then the extended teaching dimension is defined as follows.

Definition 8.4. For any nonempty sets $\mathcal{H} \subseteq \mathbb{C}$ and $\mathcal{U} \subseteq \mathcal{X}$, for any function $f : \mathcal{X} \to \mathcal{Y}$, define

$$\mathrm{XTD}(f, \mathcal{H}, \mathcal{U}) = \min \left\{ t : \min_{S \in \mathcal{U}^t} |\{h \in \mathcal{H} : h(S) = f(S)\}| \leq 1 \right\} \cup \{\infty\}$$

and define

$$\mathrm{XTD}(\mathcal{H}, \mathcal{U}) = \sup_{f : \mathcal{X} \to \mathcal{Y}} \mathrm{XTD}(f, \mathcal{H}, \mathcal{U}).$$

For a given function $f : \mathcal{X} \to \mathcal{Y}$, a minimum-sized sequence S of points in $\mathcal{U}$ such that $|\{h \in \mathcal{H} : h(S) = f(S)\}| \leq 1$ is called a *minimal specifying set* for f on $\mathcal{U}$ with respect to $\mathcal{H}$.

Note that the function f in the supremum ranges over *all* functions $\mathcal{X} \to \mathcal{Y}$, including those not contained in the hypothesis class. The extended teaching dimension has been calculated for several learning problems. For instance, if $n \in \mathbb{N}$, $z_1 \leq x_1 < z_2 \leq x_2 < \cdots < z_{n-1} \leq x_{n-1} < z_n$ are real values, $\mathcal{H} = \{\mathbb{1}^{\pm}_{[z_i, \infty)} : i \in \{1, \ldots, n\}\}$ (a set of threshold classifiers), and $\mathcal{U} = \{x_1, \ldots, x_{n-1}\}$, then $\mathrm{XTD}(\mathcal{H}, \mathcal{U}) = 2$; any f with $f(x_{n-1}) = -1$ has $\{x_{n-1}\}$ as a minimal specifying set, any f with $f(x_1) = +1$ has $\{x_1\}$ as a minimal specifying set, and any other f has some $i \in \{1, \ldots, n-2\}$ with $f(x_i) = -1$ and $f(x_{i+1}) = +1$, so that $\{x_i, x_{i+1}\}$ is a specifying set, and since any one point x_j has at least two classifiers consistent with the $f(x_j)$ label (namely, $\{\mathbb{1}^{\pm}_{[z_j, \infty)}, \mathbb{1}^{\pm}_{[z_{j-1}, \infty)}\}$ if $f(x_j) = +1$, or $\{\mathbb{1}^{\pm}_{[z_{j+1}, \infty)}, \mathbb{1}^{\pm}_{[z_{j+2}, \infty)}\}$ if $f(x_j) = -1$), $\{x_i, x_{i+1}\}$ is a minimal specifying set. Hegedüs [1995] also bounds $\mathrm{XTD}(\mathcal{H}, \mathcal{U})$ for $\mathcal{H}$ a set of linear separators. Specifically, if $n, k \in \mathbb{N}$, $\mathcal{U} = \{0, 1, \ldots, k-1\}^n$, and $\mathcal{H}$ is a subset of the class of linear separators (Example 3) such that $\forall h, g \in \mathcal{H}, \exists x \in \mathcal{U}$ with $h(x) \neq g(x)$, then $\mathrm{XTD}(\mathcal{H}, \mathcal{U}) \lesssim (\log(k))^{n-1}$ (considering n as a constant).

For any f, and any number t of queries not sufficiently large to guarantee $\min_{S \in \mathcal{X}^t} |\{h \in \mathbb{C} : h(S) = f(S)\}| \leq 1$, any algorithm making at most t queries could potentially be faced with answers consistent with f when the target $f^\star$ is chosen among the resulting multiple elements of $\mathbb{C}$ also consistent with these answers; thus, such a t is not sufficiently large to identify the target $f^\star$, in the worst case over $f^\star \in \mathbb{C}$. Hegedüs [1995] uses this reasoning to prove $\mathrm{XTD}(\mathbb{C}, \mathcal{X}) \leq \Lambda^*_{\mathrm{MQ}}(\mathbb{C})$. Furthermore, he proves a related upper bound by proposing and analyzing a simple algorithm, which essentially implements the *Halving algorithm* of Littlestone [1988], based on using minimal specifying sets to identify mistake points for a certain carefully-chosen hypothesis, thereby geometrically reducing the number of classifiers in the version space. This algorithm is formally stated as follows.

For any finite set $\mathcal{H} \subseteq \mathbb{C}$, the *majority vote classifier* of $\mathcal{H}$ is a function $h_{\mathrm{maj}} : \mathcal{X} \to \mathcal{Y}$ such that, $\forall x \in \mathcal{X}$, $|\{h \in \mathcal{H} : h(x) = h_{\mathrm{maj}}(x)\}| \geq |\{h \in \mathcal{H} : h(x) \neq h_{\mathrm{maj}}(x)\}|$, where ties are broken arbitrarily.

Algorithm: **MembHalving**
0. $V \leftarrow \mathbb{C}$
1. Repeat until $|V| = 1$
2. Let h_{maj} be the majority vote classifier of V
3. Let S be a minimal specifying set for h_{maj} on $\mathcal{X}$ w.r.t. V
4. Request the label $f^\star(x)$ for every $x \in S$
5. Let $V \leftarrow \{h \in V : h(S) = f^\star(S)\}$
6. Return the remaining element of V

By definition of XTD, each round of the loop will query for at most $\mathrm{XTD}(V, \mathcal{X}) \leq \mathrm{XTD}(\mathbb{C}, \mathcal{X})$ labels. Furthermore, the responses will either satisfy $h_{\mathrm{maj}}(S) = f^\star(S)$, in which case (by definition of a minimal specifying set) we will have $|V| = 1$ in Step 5, or else $h_{\mathrm{maj}}(S) \neq f^\star(S)$, which means at least half of the classifiers in V disagree with $f^\star$ on some point in S, so that $|V|$ will be (at least) halved in Step 5. Thus, this algorithm must satisfy the condition $|V| = 1$ to break the loop within $\log_2(|\mathbb{C}|)$ rounds. Combining this argument with the above lower-bound reasoning, along with a simple coding argument to supplement the lower bound, Hegedüs [1995] proves the following result.

Theorem 8.5.

$$\max\{\mathrm{XTD}(\mathbb{C}, \mathcal{X}), \lceil \log_2(|\mathbb{C}|) \rceil\} \leq \Lambda^*_{\mathrm{MQ}}(\mathbb{C}) \leq \mathrm{XTD}(\mathbb{C}, \mathcal{X}) \lceil \log_2(|\mathbb{C}|) \rceil.$$

This analysis is fairly tight (though Hegedüs, 1995, also shows the upper bound can be improved by a factor of $\frac{1}{2}\log_2(\mathrm{XTD}(\mathbb{C},\mathcal{X}) \vee 2)$ if Step 4 queries the points in S in a carefully-chosen order, and stops querying upon reaching a point $x \in S$ with $f^\star(x) \neq h_{\mathrm{maj}}(x)$). Interestingly, however, there are other simpler algorithms that also achieve the upper bound stated in Theorem 8.5. One such algorithm is the simple *greedy* strategy, which always queries the point guaranteed to eliminate the most surviving classifiers for its worst-case label (this is a special case of a heuristic known as *uncertainty sampling*). The algorithm is formally stated as follows.

Algorithm: **Greedy**
0. $V \leftarrow \mathbb{C}$
1. Repeat until $\lvert V\rvert = 1$
2. Let $\tilde{x} \leftarrow \mathrm{argmax}_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} \lvert\{h \in V : h(x) \neq y\}\rvert$
3. Request the label $f^\star(\tilde{x})$
4. Let $V \leftarrow \{h \in V : h(\tilde{x}) = f^\star(\tilde{x})\}$
5. Return the remaining element of V

Balcázar, Castro, and Guijarro [2002] prove that this Greedy algorithm is an MQ-algorithm for Exact learning $\mathbb{C}$, guaranteed to make a number of queries at most $\mathrm{XTD}(\mathbb{C},\mathcal{X})\lceil\ln(|\mathbb{C}|)\rceil$. The essential argument is that, on each round, if h_{maj} is the majority vote classifier of V, then since we require at most $\mathrm{XTD}(V,\mathcal{X}) \leq \mathrm{XTD}(\mathbb{C},\mathcal{X})$ points to form a minimal specifying set for h_{maj} on $\mathcal{X}$ with respect to V, there must be at least one $x \in \mathcal{X}$ with $|\{h \in V : h(x) \neq h_{\mathrm{maj}}(x)\}| \geq (|V|-1)/\mathrm{XTD}(\mathbb{C},\mathcal{X})$. Since we always have $|\{h \in V : h(x) = h_{\mathrm{maj}}(x)\}| \geq |V|/2$, this implies $\min_{y\in\mathcal{Y}} |\{h \in V : h(\tilde{x}) \neq y\}| \geq \min\{(|V|-1)/\mathrm{XTD}(\mathbb{C},\mathcal{X}), |V|/2\}$, so that $|V|$ decreases geometrically in the number of rounds; some further reasoning about the implied recurrence leads to the stated bound.

More recently, Nowak [2008, 2011] has produced an alternative analysis of the Greedy algorithm (under the name *Generalized Binary Search*), leading to a complexity measure that, though not always as small as $\mathrm{XTD}(\mathbb{C},\mathcal{X})$, is often much easier to calculate, even for interesting and broad classes of learning problems. Specifically, the following definition is equivalent to that of Nowak [2011]. Let $\mathbb{C}_-$ be a subset of $\mathbb{C}$ such that, for any $h \in \mathbb{C}$, $|\{h, -h\} \cap \mathbb{C}_-| = 1$: that is, for any $h \in \mathbb{C}$ with $-h \in \mathbb{C}$ as well, we omit one of the two classifiers when forming $\mathbb{C}_-$, but include everything else from $\mathbb{C}$. For any $k \in \mathbb{N} \cup \{0\}$, two points $x, x' \in \mathcal{X}$ are said to be *k-neighbors* if $|\{h \in \mathbb{C}_- : h(x) \neq h(x')\}| \leq k$. Now consider the set E_k of all pairs $\{x, x'\} \subseteq \mathcal{X}$ such that x and x' are k-neighbors; we will think of E_k as the set of *edges* in a graph (known as the k-neighborhood graph), and then naturally a pair of points $x, x' \in \mathcal{X}$ are said to be *connected* in the k-neighborhood graph if there exists a sequence $x_1, \ldots, x_t$ in $\mathcal{X}$ (for some $t \in \mathbb{N}$) such that $x_1 = x$, $x_t = x'$, and every $i \in \{1, \ldots, t-1\}$ has $\{x_i, x_{i+1}\} \in E_k$. Finally, $(\mathcal{X}, \mathbb{C})$ is said to be *k-neighborly* if the k-neighborhood graph is *connected*, in the sense that every pair $x, x' \in \mathcal{X}$ is connected in the k-neighborhood graph. Also define the *coherence parameter* $c^* = \min_{\mathcal{P}} \max_{h \in \mathbb{C}} |\int h \mathrm{d}\mathcal{P}|$, which is effectively a measure of the *balancedness* of classifiers in $\mathbb{C}$ under a most-favorable distribution $\mathcal{P}$; values of c^* close to 0 are considered favorable.

One can show that several interesting classes satisfy the k-neighborly condition, with reasonable values of k and c^*. For instance, if $\mathcal{X} = \mathbb{R}$ and $\mathbb{C} = \{\mathbb{1}^{\pm}_{[z_i, \infty)} : i \in \{1, \ldots, n\}\}$ is any finite set of *threshold* classifiers, where $\{z_i\}_{i=1}^n$ is an increasing sequence of values in $\mathbb{R}$, and $n \in \mathbb{N}$, then $(\mathcal{X}, \mathbb{C})$ is 1-neighborly; to see this, consider the partition $(-\infty, z_1)$, $[z_1, z_2)$, $\ldots$, $[z_n, \infty)$, and note that any points in adjacent regions of this partition are 1-neighbors. Furthermore, in this case we have $c^* = 0$, obtained by any $\mathcal{P}$ with $\mathcal{P}((-\infty, z_1)) = \mathcal{P}((z_n, \infty)) = 1/2$. A more interesting example from Nowak [2011] is given by the case of $\mathcal{X} = \mathbb{R}^n$ for $n \in \mathbb{N}$, and $\mathbb{C}$ an arbitrary finite set of distinct *linear separators*; Nowak [2011] proves that $(\mathcal{X}, \mathbb{C})$ is 1-neighborly in this scenario as well. The reasoning is essentially the same as for the thresholds example above (i.e., points in adjacent regions of the join of the positive regions of classifiers in $\mathbb{C}$ are 1-neighbors). He further shows that $c^* = 0$, approached by taking $\mathcal{P}$ uniform on the surface of a sphere of radius r; as $r \to \infty$, every h in the finite set $\mathbb{C}$ becomes balanced (i.e., $|\int h \mathrm{d}\mathcal{P}| \to 0$).

Nowak [2011] shows that, if $(\mathcal{X}, \mathbb{C})$ is k-neighborly with coherence parameter c^*, then the above Greedy algorithm makes a number of queries at most

$$\left\lceil \max\left\{k+1, \frac{2}{1-c^*}\right\} \ln(|\mathbb{C}|)\right\rceil$$

(Nowak's original bound is sometimes slightly smaller, but is within a factor of 2 of this). The essential idea is found by observing that the distribution $\mathcal{P}$ obtaining the min in the definition of c^* has $c^* \geq \frac{1}{|V|}\sum_{h\in V}|\int h\mathrm{d}\mathcal{P}| \geq \frac{1}{|V|}|\int(\sum_{h\in V} h)\mathrm{d}\mathcal{P}|$ so that $|\int(\sum_{h\in V} h)\mathrm{d}\mathcal{P}| \leq c^*|V|$. For any $c \geq c^*$, if it happens that every x has $|\sum_{h\in V} h(x)| > c|V|$, then the value of $\sum_{h\in V} h(x)$ must be positive for some points x and negative for others, to maintain $|\int(\sum_{h\in V} h)\mathrm{d}\mathcal{P}| \leq c^*|V|$. In particular, by the k-neighborly property, there should exist some pair $\{z, z'\}$ of k-neighbors with sums of opposite signs, so that $\sum_{h\in V} h(z)$ and $\sum_{h\in V} h(z')$ differ by more than $2c|V|$. But we know that $\sum_{h\in V} h(z)$ and $\sum_{h\in V} h(z')$ are within $2k$ of each other, so we must have $2k > 2c|V|$, and therefore $|V| < k/c$. In other words, either Greedy can make significant progress in reducing $|V|$ (i.e., $\exists x$ s.t. $|\sum_{h\in V} h(x)| \leq c|V|$), or else $|V|$ is already relatively small (i.e., $|V| < k/c$) so that only a few additional rounds are needed anyway.

Interestingly, the k-neighborly property can also be quite useful for bounding $\mathrm{XTD}(\mathbb{C}, \mathcal{X})$, as reflected in the following result.

Lemma 8.6. For any $k \in \mathbb{N}$, if $(\mathbb{C}, \mathcal{X})$ is k-neighborly, then $\mathrm{XTD}(\mathbb{C}, \mathcal{X}) \leq \max\left\{k+1, \mathrm{XTD}(\mathbb{1}^{\pm}_{\{\}}, \mathbb{C}, \mathcal{X}), \mathrm{XTD}(\mathbb{1}^{\pm}_{\mathcal{X}}, \mathbb{C}, \mathcal{X})\right\}$.

Proof. It suffices to show that any $f : \mathcal{X} \to \mathcal{Y}$ with $\min_{x\in\mathcal{X}} f(x) = -1$ and $\max_{x\in\mathcal{X}} f(x) = +1$ has $\mathrm{XTD}(f, \mathbb{C}, \mathcal{X}) \leq k+1$. Fix such a function f. Let $x, x' \in \mathcal{X}$ be such that $f(x) \neq f(x')$. Since $(\mathbb{C}, \mathcal{X})$ is k-neighborly, there is a finite-length path in the k-neighborhood graph connecting x and x'. There must be two k-neighbors $z, z' \in \mathcal{X}$ along this path such that $f(z) = f(x) \neq f(x') = f(z')$. By the k-neighbor property, there are at most k elements $h \in \mathbb{C}$ with $h(z) = f(z)$ and $h(z') = f(z')$ (noting that, for functions h with $-h \in \mathbb{C}$, at most one of the two will agree with f on these points). Say k' such functions exist, for some $k' \in \{0, 1, \ldots, k\}$. Since $\mathbb{C}$ is defined as a set of functions on $\mathcal{X}$, each element has a distinct classification of $\mathcal{X}$, so that at most one of these k' classifiers agrees with f on all of $\mathcal{X}$. Therefore, for (at least) $k' - 1$

of these classifiers, there exists a point on which it disagrees with f. Combining $k' - 1$ of these points with z and z', we have a specifying set for f on $\mathcal{X}$ with respect to $\mathbb{C}$, of size $k' + 1 \leq k + 1$. $\square$

8.3.2 Extension to Statistical Learning

Section 8.3.1 described a characterization of the number of queries required for Exact learning with membership queries, expressed in terms of the extended teaching dimension. It turns out the extended teaching dimension is also useful for studying the label complexity of active learning in the statistical learning setting studied in the present article.

For any $m \in \mathbb{N}$ and $\mathcal{U} \in \mathcal{X}^m$, and any $\mathcal{H} \subseteq \mathbb{C}$, let $\mathcal{H}[\mathcal{U}]$ denote a maximal subset of $\mathcal{H}$ such that $\forall h, g \in \mathcal{H}, h(\mathcal{U}) \neq g(\mathcal{U})$: that is, $\mathcal{H}[\mathcal{U}]$ contains exactly one classifier from $\mathcal{H}$ for each labeling of $\mathcal{U}$ realized by classifiers in $\mathcal{H}$. Also, for $f : \mathcal{X} \to \mathcal{Y}$, we overload the notation $\mathrm{XTD}(f, \mathcal{H}, \mathcal{U})$ and $\mathrm{XTD}(\mathcal{H}, \mathcal{U})$ to allow $\mathcal{U}$ to be a sequence (rather than a set) by taking the set of distinct entries in $\mathcal{U}$.

Given these definitions, one obvious way to use the extended teaching dimension (in the realizable case) is to apply the MembHalving algorithm to the set $\mathbb{C}[\mathcal{U}_m]$, where $\mathcal{U}_m = \{X_1, \ldots, X_m\}$, and m is a sufficiently large integer. In the realizable case, this will be guaranteed to identify $Y_1, \ldots, Y_m$, so that we can then use any passive learning algorithm on the data set $\mathcal{Z}_m$. By the above analysis, the number of labels requested by this method would be at most $\mathrm{XTD}(\mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)\lceil \log_2(|\mathbb{C}[\mathcal{U}_m]|)\rceil$. Since it is known that $\log_2(|\mathbb{C}[\mathcal{U}_m]|) \leq d \log_2(em/d)$ [Vapnik and Chervonenkis, 1971], combined with Theorem 3.2 (to identify a sufficient size for m), the label complexity of this method is $\lesssim (\sup_{\mathcal{U} \in \mathcal{X}^m} \mathrm{XTD}(\mathbb{C}[\mathcal{U}], \mathcal{U}))\, d\mathrm{Log}\,(em/d)$, where $m \lesssim (1/\varepsilon)(d\mathrm{Log}(1/\varepsilon) + \mathrm{Log}(1/\delta))$. However, Hanneke [2007a] found that it is possible to refine this analysis by reducing the sizes of the sets $\mathcal{U}$ the specifying sets are constructed for in MembHalving, and also incorporating the distribution $\mathcal{P}$ in the analysis. The key insight is that we only need the unlabeled sample to be large enough so that, if we do not find that the majority vote classifier makes a mistake on that set, then we can be confident that the majority vote classifier has low error rate, and therefore can be returned. This reasoning motivates the following algorithm due to Hanneke [2007a] (slightly modified here to match our present budget-based active learning setting).

Algorithm: **ActiveHalving**$_{m,\delta}(n)$
0. Let V_0 be a minimal $(\delta/(2mn))$-cover of $\mathbb{C}$; $t \leftarrow 0$, $i \leftarrow 0$
1. Repeat
2. Let h_i be the majority vote classifier of V_i
3. Let $\mathcal{U}^{(i)} = \{X_{im+1}, \ldots, X_{(i+1)m}\}$
4. Let S_i be a minimal specifying set for h_i on $\mathcal{U}^{(i)}$ w.r.t. $V_i[\mathcal{U}^{(i)}]$
5. If $t + |S_i| \vee 1 \leq n$
6. Request the label Y_j for every $X_j \in S_i$; let $t \leftarrow t + |S_i| \vee 1$
7. Let $V_{i+1} \leftarrow \{h \in V_i : \forall X_j \in S_i, h(X_j) = Y_j\}$; $i \leftarrow i + 1$
8. Else Return $h_{\hat{i}}$, where
$\hat{i} = \operatorname{argmin}_{i' < i \vee 1} \min_{h \in V_{i'+1}} \sum_{k=i'm+1}^{(i'+1)m} |h(X_k) - h_{i'}(X_k)|$

To characterize the label complexity of this method, Hanneke [2007a] proposes the following quantity.

Definition 8.7. For any $m \in \mathbb{N}$ and $\delta \in (0, 1)$, for $\mathcal{U}_m = \{X_1, \ldots, X_m\}$, define

$$\mathrm{XTD}(m, \delta) = \min \{t : \forall f, \mathbb{P}(\mathrm{XTD}(f, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m) > t) \leq \delta\},$$

where f ranges over all classifiers.

With this definition, we can express a bound on the label complexity of the ActiveHalving algorithm as follows [Hanneke, 2007a].

Theorem 8.8. For any $\varepsilon, \delta \in (0, 1)$, letting $\delta' = \frac{\delta}{24d \log_2\left(\frac{4d}{\varepsilon\delta}\right)}$, for any $m \geq \left\lceil \frac{4}{\varepsilon} \mathrm{Log}\left(\frac{1}{\delta'}\right) \right\rceil$, ActiveHalving$_{m,\delta}$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ in the realizable case,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \mathrm{XTD}(m, \delta') d \mathrm{Log}\left(\frac{dm}{\delta}\right).$$

The quantity $\mathrm{XTD}(m, \delta)$ has been bounded for a few different types of learning problems. For the problem of learning threshold classifiers (Example 1), we clearly have $\mathrm{XTD}(m, \delta) \leq 2$, by essentially the same reasoning given above (after Definition 8.4). Hanneke [2007a] additionally bounds $\mathrm{XTD}(m, \delta)$ for the problem of learning not-too-small *axis-aligned rectangles* over $\mathbb{R}^k$, when $\mathcal{P}$ is a product distribution with a density function (with respect to the Lebesgue measure). Specifically, in this learning problem, the hypothesis class is specified as $\mathbb{C} = \{\mathbb{1}^{\pm}_{\times_{i=1}^{k}[z_{2i-1}, z_{2i}]} : z_1, \ldots, z_{2k} \in \mathbb{R}, \mathcal{P}(\times_{i=1}^{k}[z_{2i-1}, z_{2i}]) \geq p\}$, for some value $p \in (0, 1)$. In this case, Hanneke [2007a] finds that $\mathrm{XTD}(m, \delta) \lesssim \frac{k^2}{p} \mathrm{Log}\left(\frac{km}{\delta}\right)$.

8.3.3 Robustness to Noise

The ideas leading to the ActiveHalving algorithm also generalize to noisy settings. The challenge in noisy settings is that the samples $\mathcal{U}^{(i)}$ will often contain points X_j with $Y_j \neq f^{\star}(X_j)$, and if one of these points is included in the specifying set S_i, then $h^* = \mathrm{argmin}_{h \in V_i} \mathrm{er}(h)$ may be inconsistent with the responses to the label requests; thus, we cannot simply remove a classifier from V_i after making a single mistake, without risking discarding h^*. The main trick explored by Hanneke [2007a] to compensate for this is to take many small subsamples of size $\lesssim \frac{1}{\nu+\varepsilon}$. Since most subsamples of this size will not contain any points X_j with $Y_j \neq h^*(X_j)$, h^* is not likely to be contradicted by the labels of more than a small fraction of the corresponding minimal specifying sets; thus, we can confidently discard any classifier in V_i contradicted on a large fraction of these sets. As long as the majority vote classifier has large error rate, there will be many classifiers making this large number of mistakes, and thus $|V_i|$ will decrease geometrically.

However, this trick only works as long as the majority vote classifier h_{maj} has $\mathrm{er}(h_{\mathrm{maj}}) > c(\nu + \varepsilon)$ for a constant $c > 1$. Once we have made enough progress that this fails to hold, h_{maj} is likely to *agree* with h^* on many of these samples, so that there may no longer be many classifiers making enough mistakes to determine they are not h^*. Therefore, to choose a classifier to return from among these remaining classifiers, we use the samples of size $\lesssim \frac{1}{\nu+\varepsilon}$ in a slightly different way; in fact, we *use* the fact that these samples $\mathcal{U}^{(i)}$ will frequently have $h_{\mathrm{maj}}(\mathcal{U}^{(i)}) = h^*(\mathcal{U}^{(i)})$. Specifically, we take a large number of unlabeled samples, and sample many subsets, again of size $\lesssim \frac{1}{\nu+\varepsilon}$, and request the labels for a minimal specifying set (for h_{maj}) for each of these subsamples. By forming enough samples of this type, each of these unlabeled points appears in many such subsamples, and we can thereby ensure that each of these data points either has its label directly requested, or else it appears in a larger number of samples $\mathcal{U}^{(i)}$ for which both h_{maj} and h^* are consistent with the labels of the minimal specifying set, compared to the number of samples $\mathcal{U}^{(i)}$ it appears in for which one of these classifiers makes a mistake on the minimal specifying set. Anytime h_{maj} is consistent with the labels of a specifying set (for h_{maj} on $\mathcal{U}^{(i)}$ with respect to $V[\mathcal{U}^{(i)}]$), there is at most one consistent classification of $\mathcal{U}^{(i)}$ consistent with a classifier in V, and if h^* is also consistent with these labels, this one consistent classification must be $h^*(\mathcal{U}^{(i)})$. Therefore, for each unqueried unlabeled sample X_j, we can simply take a vote of the consistent classification of X_j over all such occurrences for subsamples $\mathcal{U}^{(i)}$ with $X_j \in \mathcal{U}^{(i)}$, and thereby *infer* the value $h^*(X_j)$. We thus arrive at a labeling of the collection of unlabeled samples with the property that each label is either the Y_i value (for queried samples) or the $h^*(X_i)$ value (for unqueried samples). By feeding this labeled data set (of an appropriately large size) into an empirical risk minimization algorithm, we arrive at a classifier $\hat{h}$ with $\mathrm{er}(\hat{h}) \leq \nu + \varepsilon$.

The above reasoning is formalized by Hanneke [2007a] to arrive at a noise-robust active learning algorithm, called ReduceAndLabel. That work further shows that, for any $\mathcal{P}_{XY}$, for any $\varepsilon, \delta \in (0,1)$, the ReduceAndLabel algorithm achieves a label complexity Λ such that

$$\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \mathrm{XTD}(u, \delta') \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right),$$

where $u = \left\lfloor \frac{1}{16(\nu+3\varepsilon/4)} \right\rfloor$ and $\delta' = \mathrm{poly}(\varepsilon\delta/d)$. However, one significant catch with this algorithm is that, in order for the sample sizes to be set appropriately, we need direct access to the noise rate ν (though Hanneke, 2007a, shows it also suffices to have a reasonably good upper bound on ν); it is not presently known whether this limitation can be removed while still satisfying the above label complexity bound, for instance by a method that adaptively chooses the sample sizes.

Balcan and Hanneke [2012] have recently applied this same technique in a more general setting, allowing general abstract families of queries, beyond label requests: for instance, queries that propose a set $\mathcal{U} \subseteq \{X_1, \ldots, X_m\}$ and a label $y \in \mathcal{Y}$ and ask that an example $X_i \in \mathcal{U}$ with $Y_i = y$ be returned if one exists; they further study this for the general case of $|\mathcal{Y}| \in \mathbb{N}$, rather than merely for binary classification. They generalize $\mathrm{XTD}(m, \delta)$ to an abstract quantity suitable for these other families of queries, based on a related generalization of Balcázar, Castro, and Guijarro [2002] from the Exact learning setting. Interestingly, they find that some types of queries allow one to significantly simplify the second phase of the algorithm (constructing the labeled data set), and can also be used to adapt to the value of ν so that no direct information about $\mathcal{P}_{XY}$ is required. For some types of queries, they also extend this technique to the problem of learning with bounded noise (i.e., satisfying (2.2) for some value $a \in (1, \infty)$), where they find that it is sometimes possible to achieve error rate less than $\nu + \varepsilon$ using a number of queries smaller than the number of label requests required by certain types of active learning algorithms, improving the dependence on a by roughly a factor of $\frac{2a}{a-1}$. They further show this is the best one can hope to achieve with these types of queries: that is, the required number of queries would be at least $\frac{a-1}{2a}$ times the label complexity of active learning.

8.4 An Alternative Analysis of CAL

There is an interesting topic in the machine learning literature, known as *selective classification*, which shares many fundamental features with the active learning problem. The objective in selective classification is to produce a pair of functions (f, g), where f is a classifier and g is a measurable function mapping $\mathcal{X}$ to $[0, 1]$. We interpret this as saying that, for any point x, the selective classifier will predict $f(x)$ with probability $g(x)$, and otherwise it will simply *refuse* to make a prediction. The performance is then measured by *both* $R(f, g) = \mathbb{E}[\mathbb{1}[f(X) \neq Y]g(X)]/\mathbb{E}[g(X)]$ (the probability it makes a mistake given that it makes a prediction, called *risk*) *and*

$\Phi(f,g) = \mathbb{E}[g(X)]$ (the probability it makes a prediction, called *coverage*), where $(X,Y) \sim \mathcal{P}_{XY}$. Thus, there is a trade-off between risk and coverage when designing and analyzing a selective classification algorithm. For completeness, define $R(f,g) = 0$ when $\Phi(f,g) = 0$.

El-Yaniv and Wiener [2010] study an extreme form of selective classification, which they call *perfect* selective classification, in which the algorithm is required to produce a pair (f,g) with the property that, for every $\mathcal{P}_{XY}$ in the realizable case, $R(f,g) = 0$ (with certainty): that is, whenever the selective classifier makes a prediction, it is always correct. They find that a slight modification of CAL leads to a perfect selective classification algorithm (which they call *consistent selective strategy*, or CSS), and in fact that it achieves the maximum possible coverage among all perfect selective classification strategies. Specifically, the CSS algorithm, applied to $\mathcal{Z}_m$, simply takes f as any element of the version space $V_m^\star$, and takes $g = \mathbb{1}_{\mathcal{X}\setminus\mathrm{DIS}(V_m^\star)}$. This is clearly a perfect selective classification algorithm, since any x on which $f(x) \neq f^\star(x)$ will have $x \in \mathrm{DIS}(V_m^\star)$, and therefore $g(x) = 0$. The coverage of CSS is therefore precisely $1 - \mathcal{P}(\mathrm{DIS}(V_m^\star))$. Thus, the analysis of perfect selective classification is largely concerned with bounding $\mathcal{P}(\mathrm{DIS}(V_m^\star))$.

Since the analysis of the label complexity of CAL is also concerned with the region $\mathrm{DIS}(V_m^\star)$, it is natural to suspect that any analysis of the coverage of CSS should translate into a bound on the label complexity of CAL. Indeed, by a slight modification of the proof of Theorem 5.1, El-Yaniv and Wiener [2012] make precisely this connection. Specifically, suppose we have a function $\Delta : \mathbb{N} \times (0,1) \to [0,1]$ with the property that, for any $m \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$, $\mathcal{P}(\mathrm{DIS}(V_m^\star)) \leq \Delta(m,\delta)$. Following the proof of Theorem 5.1 and using the notation introduced there, monotonicity of $m \mapsto \mathcal{P}(\mathrm{DIS}(V_{m-1}^\star))$ and a union bound imply that with probability at least $1 - \sum_{i=0}^{i_\varepsilon - 1} \frac{\delta}{3(2+i_\varepsilon - i)^2} > 1 - \delta/3$,

$$\sum_{m=1}^{m_{i_\varepsilon}} \mathcal{P}(\mathrm{DIS}(V_{m-1}^\star)) \leq \sum_{i=1}^{i_\varepsilon} (m_i - m_{i-1}) \mathcal{P}(\mathrm{DIS}(V_{m_{i-1}}^\star))$$

$$\leq \sum_{i=1}^{i_\varepsilon} m_i \Delta\left(m_{i-1}, \frac{\delta}{3(3+i_\varepsilon - i)^2}\right). \quad (8.4)$$

Noting that $\{\mathbb{1}_{\mathrm{DIS}(V^\star_{m-1})}(X_m) - \mathcal{P}(\mathrm{DIS}(V^\star_{m-1}))\}_{m=1}^\infty$ is a martingale difference sequence with respect to $\{X_m\}_{m=1}^\infty$, Bernstein's inequality (for martingales) implies that with probability at least $1-\delta/3$, if (8.4) holds then

$$\sum_{m=1}^{m_{i_\varepsilon}} \mathbb{1}_{\mathrm{DIS}(V^\star_{m-1})}(X_m) \lesssim \mathrm{Log}(1/\delta) + \sum_{i=1}^{i_\varepsilon} m_i \Delta\left(m_{i-1}, \frac{\delta}{3(3+i_\varepsilon - i)^2}\right)$$

[e.g., van de Geer, 2000, El-Yaniv and Wiener, 2012]. Combining these two facts with Lemma 3.1, (3.2), and a union bound, we see that CAL achieves a label complexity Λ such that, for $\mathcal{P}_{XY}$ in the realizable case, $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \mathrm{Log}\,(1/\delta) + \sum_{i=1}^{\lceil \log_2(1/\varepsilon) \rceil} m_i \Delta\left(m_{i-1}, \frac{\delta}{3(\lceil \log_2(8/\varepsilon) \rceil - i)^2}\right). \tag{8.5}$$

This abstract label complexity bound was originally obtained by El-Yaniv and Wiener [2012] (with minor differences).

Note that, as we did in the proof of Theorem 5.1 in Chapter 5, one can easily use the disagreement coefficient to express a function $\Delta(m, \delta)$ with the above property, since Lemma 3.1 implies that with probability at least $1-\delta$, $\mathcal{P}(\mathrm{DIS}(V^\star_m)) \le \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, U(m,\delta)))) \le \theta(U(m,\delta))U(m,\delta)$. Indeed, plugging this last quantity into (8.5) for $\Delta(\cdot,\cdot)$, and following the original proof of Theorem 5.1 to simplify the expression, one obtains precisely the label complexity bound of Theorem 5.1 (up to constant factors). Furthermore, in light of Theorem 5.2, defining $\Delta(m,\delta)$ in this way should often be relatively tight. Specifically, a straightforward combination of Theorem 5.2 and Markov's inequality reveals that any $\Delta(\cdot,\cdot)$ with the above property has

$$\Delta(m, \delta) \ge (1/7)\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, 1/m))) \tag{8.6}$$

for any $\delta \in (0, 1/8)$ and $m \ge 2$ [El-Yaniv and Wiener, 2012]; in particular, this implies $\Delta(m,\delta) \ne o(\theta(1/m)/m)$.

However, interestingly, El-Yaniv and Wiener [2010] identify an entirely novel way to bound the coverage of CSS, and therefore the label complexity of CAL, in terms of a new complexity measure that has several noteworthy features. It incorporates certain aspects of many different known complexity measures, including the notion of a region of disagreement (as in the disagreement coefficient analysis), the notion of a minimal specifying set (as in the teaching dimension analysis of Section 8.3), and the notion of the VC dimension. The specific quantity can be summarized as the VC dimension of the set of regions of disagreement of version spaces that can be arrived at by observing a number of points equal the size of a minimal specifying set for $f^\star$ on $\{X_1, \ldots, X_m\}$ with respect to $\mathbb{C}[\{X_1, \ldots, X_m\}]$. This is formalized in the following definition, due to El-Yaniv and Wiener [2010].

For $\mathcal{U}_m = \{X_1, \ldots, X_m\}$, let $\hat{n}_m = \mathrm{XTD}(f^\star, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$ be the size of a minimal specifying set for $f^\star$ on $\mathcal{U}_m$ with respect to $\mathbb{C}[\mathcal{U}_m]$; El-Yaniv and Wiener [2010, 2012] refer to $\hat{n}_m$ as the *version space compression set size*. For any $n \in \mathbb{N}$ and $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^n$, let $\mathbb{C}[\mathcal{L}] = \{h \in \mathbb{C} : \mathrm{er}_{\mathcal{L}}(h) = 0\}$, and define $\mathbb{D}_n = \{\mathbb{1}^{\pm}_{\mathrm{DIS}(\mathbb{C}[\mathcal{L}])} : \mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^n\}$ and $\gamma(\mathbb{C}, n) = \mathrm{vc}\,(\mathbb{D}_n)$, the VC dimension of $\mathbb{D}_n$, referred to as the *order-n characterizing set complexity* of $\mathbb{C}$.

For reasons explained below, we are particularly interested in the value $\gamma(\mathbb{C}, \hat{n}_m)$. Note that $\hat{n}_m$ depends on the data $\mathcal{Z}_{\mathbf{X}}$ itself, so that $\gamma(\mathbb{C}, \hat{n}_m)$ is a random variable. However, for several learning problems, El-Yaniv and Wiener [2010, 2012] obtain interesting data-independent upper bounds for it. For instance, for the problem of learning threshold classifiers (Example 1), $\hat{n}_m \leq 2$ (as in Section 8.3.1), and since, for any $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^{\hat{n}_m}$, the region $\mathrm{DIS}(\mathbb{C}[\mathcal{L}])$ is either empty or an *interval*, $\gamma(\mathbb{C}, \hat{n}_m) \leq 2$. For the problem of learning interval classifiers (Example 2), for any $m \geq \frac{1}{\mathcal{P}(x: f^\star(x) = +1)} \mathrm{Log}(1/\delta)$, with probability at least $1 - \delta$, at least one $i \leq m$ has $f^\star(X_i) = +1$, so that $\hat{n}_m \leq 4$ (taking the ≤ 2 points adjacent to each boundary); any $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^{\hat{n}_m}$ has $\mathrm{DIS}(\mathbb{C}[\mathcal{L}])$ either empty, a union of two intervals, or a set $\mathcal{X} \setminus \{x_1, \ldots, x_{\hat{n}_m}\}$ for some points $x_1, \ldots, x_{\hat{n}_m} \in \mathcal{X}$ (for the case where the points in $\mathcal{L}$ are all labeled negative). Thus, in the case of $\hat{n}_m \leq 4$, we have $\gamma(\mathbb{C}, \hat{n}_m) \leq 4$; however, if no $i \leq m$ has $f^\star(X_i) = +1$ (e.g., if m is small), then $\hat{n}_m = m$, and because of the sets $\mathrm{DIS}(\mathbb{C}[\mathcal{L}])$ of type $\mathcal{X} \setminus \{x_1, \ldots, x_{\hat{n}_m}\}$, we have $\gamma(\mathbb{C}, \hat{n}_m) = m$ in this case.

El-Yaniv and Wiener [2010, 2012] also bound $\gamma(\mathbb{C}, \hat{n}_m)$ for more involved examples. For the class of k-dimensional linear separators

(Example 3), when $\mathcal{P}$ is a mixture of a finite number of multivariate normal distributions with diagonal covariance matrices of full rank, El-Yaniv and Wiener [2010] find that, with probability at least $1-\delta$, $\hat{n}_m = O\left((\mathrm{Log}(m))^{k-1}/\delta\right)$ (considering k as a constant), and in this case $\gamma(\mathbb{C}, \hat{n}_m) = O\left(\hat{n}_m^{\lfloor (k+1)/2 \rfloor}\mathrm{Log}(\hat{n}_m)\right) \leq O\left((\log(m))^{(k-1)\lfloor (k+1)/2 \rfloor}\delta^{-\lfloor (k+1)/2 \rfloor}\mathrm{Log}\left(\mathrm{Log}(m)/\delta\right)\right)$. Additionally, consider the class of not-too-small axis-aligned rectangles over $\mathbb{R}^k$ (recall the definition from Section 8.3.2), when $\mathcal{P}$ is a product distribution with a density function, and $p = \inf_{h\in\mathbb{C}} \mathcal{P}(x : h(x) = +1) > 0$. Recall from Section 8.3.2 that in this case, Hanneke [2007a] showed $\mathrm{XTD}(m,\delta) \lesssim \frac{k^2}{p}\mathrm{Log}\left(\frac{km}{\delta}\right)$. This provides a bound for $\hat{n}_m$, since we always have that with probability at least $1-\delta$, $\hat{n}_m = \mathrm{XTD}(f^\star, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m) \leq \mathrm{XTD}(m,\delta)$. Furthermore, by noting that the sets $\mathcal{X} \setminus \mathrm{DIS}(\mathbb{C}[\mathcal{L}])$ (for $\mathcal{L} \in (\mathcal{X}\times\mathcal{Y})^{\hat{n}_m}$) are representable as unions of at most $\hat{n}_m$ rectangles, El-Yaniv and Wiener [2012] show that $\gamma(\mathbb{C}, \hat{n}_m) \lesssim \frac{k^3}{p}\mathrm{Log}\left(\frac{km}{\delta}\right)$ in this case.

By a clever application of Lemma 3.1, El-Yaniv and Wiener [2010] show that, for any $\delta \in (0,1)$, $m \in \mathbb{N}$, and any sequence $p_1, \ldots, p_m \in [0,1]$ with $\sum_{i=1}^m p_i \leq 1$, with probability at least $1-\delta$,

$$\mathcal{P}(\mathrm{DIS}(V_m^\star)) \leq \min_{n\in\{\hat{n}_m,\ldots,m\}} \frac{c}{m}\left(\gamma(\mathbb{C},n)\mathrm{Log}\left(\frac{m}{\gamma(\mathbb{C},n)}\right) + \mathrm{Log}\left(\frac{1}{p_n\delta}\right)\right). \tag{8.7}$$

The reasoning is that, since we know $\frac{1}{m}\sum_{i=1}^m \mathbb{1}[\mathbb{1}^{\pm}_{\mathrm{DIS}(V_m^\star)}(X_i) \neq \mathbb{1}^{\pm}_{\{\}}(X_i)] = 0$, applying Lemma 3.1 with hypothesis class $\mathbb{D}_n$ in the realizable case with marginal distribution $\mathcal{P}$ over $\mathcal{X}$ and target function $\mathbb{1}^{\pm}_{\{\}}$ gives that with probability at least $1-p_n\delta$, if $\mathbb{1}^{\pm}_{\mathrm{DIS}(V_m^\star)} \in \mathbb{D}_n$, then $\mathcal{P}(x : \mathbb{1}^{\pm}_{\mathrm{DIS}(V_m^\star)}(x) \neq \mathbb{1}^{\pm}_{\{\}}(x)) \leq cm^{-1}(\mathrm{vc}(\mathbb{D}_n)\mathrm{Log}(m/\mathrm{vc}(\mathbb{D}_n)) + \mathrm{Log}(1/p_n\delta))$. The inequality (8.7) then follows by a union bound over values of $n \in \{1,\ldots,m\}$, since $\mathcal{P}(x : \mathbb{1}^{\pm}_{\mathrm{DIS}(V_m^\star)}(x) \neq \mathbb{1}^{\pm}_{\{\}}(x)) = \mathcal{P}(\mathrm{DIS}(V_m^\star))$, $\mathrm{vc}(\mathbb{D}_n) = \gamma(\mathbb{C},n)$, and we know $\mathbb{1}^{\pm}_{\mathrm{DIS}(V_m^\star)} \in \mathbb{D}_n$ for every $n \geq \hat{n}_m$.

In particular, suppose $n_{m,\delta}$ is an integer such that, with probability at least $1-\delta/2$, $\hat{n}_m \leq n_{m,\delta}$. In this case, taking

$$\Delta(m,\delta) = \frac{c}{m}\left(\gamma(\mathbb{C}, n_{m,\delta})\mathrm{Log}\left(\frac{m}{\gamma(\mathbb{C}, n_{m,\delta})}\right) + \mathrm{Log}\left(\frac{2}{\delta}\right)\right), \tag{8.8}$$

by setting $p_{n_{m,\delta}} = 1$, and $p_n = 0$ for all $n \neq n_{m,\delta}$, (8.7) and a union bound imply that, with probability at least $1-\delta$, $\mathcal{P}(\mathrm{DIS}(V_m^{\star})) \leq \Delta(m,\delta)$; thus, this is a valid specification of $\Delta(m,\delta)$, which can therefore be used in (8.5) to bound the label complexity of CAL [El-Yaniv and Wiener, 2012].

Comparing this technique based on $\gamma(\mathbb{C}, \hat{n}_m)$ with the analysis in terms of the disagreement coefficient above, from a practical perspective, it seems some problems are easier to approach with one or the other of these techniques. As we have seen, the process of bounding $\theta(\varepsilon)$ often focuses on determining the *volumes* of various regions of $\mathcal{X}$; on the other hand, the process of bounding $\gamma(\mathbb{C}, \hat{n}_m)$ seems to focus more on describing the *shapes* of various regions. Since there are some problems for which these shapes may be relatively easier to characterize, we might expect $\gamma(\mathbb{C}, \hat{n}_m)$ to be quite useful sometimes. For instance, at present this is the only technique known to establish the bounds on the label complexity of CAL that result from the above bounds on $\gamma(\mathbb{C}, \hat{n}_m)$ for both k-dimensional linear separators under mixtures of multivariate normal distributions and axis-aligned rectangles under product distributions.

Furthermore, used in combination with (8.6), this technique can also help to bound the disagreement coefficient itself, thus formally relating these two quantities. For instance, plugging the specification of $\Delta(m,\delta)$ from (8.8) into (8.6) and taking $\delta \in (1/16, 1/8)$, we have that for any $m \geq 2$,

$$\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^{\star}, 1/m))) \lesssim \frac{\gamma(\mathbb{C}, n_{m,1/8})}{m}\mathrm{Log}\left(\frac{m}{\gamma(\mathbb{C}, n_{m,1/8})}\right).$$

In particular, this implies that $\forall \varepsilon \in (0,1]$,

$$\theta(\varepsilon) \lesssim \max_{1 \leq m \leq 1/\varepsilon} \gamma(\mathbb{C}, n_{m,1/8})\mathrm{Log}\left(\frac{m}{\gamma(\mathbb{C}, n_{m,1/8})}\right).$$

Thus, for k-dimensional linear separators under a finite mixture of multivariate normal distributions with diagonal covariance matrices of full rank,

$$\theta(\varepsilon) = O\left((\mathrm{Log}(1/\varepsilon))^{(k-1)\lfloor (k+1)/2 \rfloor + 1}\mathrm{Log}(\mathrm{Log}(1/\varepsilon))\right)$$
$$\leq O\left((\mathrm{Log}(1/\varepsilon))^{(k^2+1)/2}\mathrm{Log}(\mathrm{Log}(1/\varepsilon))\right).$$

Interestingly, plugging this into Theorem 5.1 provides a much better dependence on δ (at the expense of a slightly worse dependence on ε) compared to the label complexity bound one finds by simply plugging the Δ values from (8.8) into (8.5) with the above bound on $\gamma(\mathbb{C}, n_{m,\delta})$. Similarly, for not-too-small axis-aligned rectangles, with $\mathcal{P}$ a product distribution with a density, and $p = \inf_{h \in \mathbb{C}} \mathcal{P}(x : h(x) = +1) > 0$,

$$\theta(\varepsilon) \lesssim \frac{k^3}{p}\mathrm{Log}\left(\frac{k}{\varepsilon}\right)\mathrm{Log}\left(\frac{p}{k\varepsilon}\right) = O\left((\mathrm{Log}(1/\varepsilon))^2\right).$$

8.5 From Disagreement to Shatterability

In some sense, disagreement-based active learning represents a kind of *baseline* for reasonable active learning methods, since it never requests a label that would definitely provide no additional information relevant to the task at hand, but does not otherwise discriminate about which labels it requests. However, as we have seen above, this approach can sometimes lead to label complexities no better than those of a comparable passive learning method; specifically, this is the case when $\theta(\varepsilon) = \Omega(1/\varepsilon)$. It is therefore natural to ask whether there are techniques that enable improvements in label complexity compared to passive learning, even when $\theta(\varepsilon) = \Omega(1/\varepsilon)$. Hanneke [2012] provides one approach to achieving this, by generalizing the notion of *disagreement* to *shatterability*: that is, in the context of CAL or RobustCAL, if we think of $\mathrm{DIS}(V)$ as the set of points x for which V shatters $\{x\}$, we can generalize this by considering a method that requests the label Y_m of X_m if V shatters $S \cup \{X_m\}$ for a carefully-chosen collection of points $S \in \mathcal{X}^k$, for some well-chosen $k \in \mathbb{N}$.

This generalization is motivated by Hanneke [2012] as follows. As mentioned, the problem with disagreement-based methods is that they do not offer improvements in label complexity compared to passive learning when $\theta(\varepsilon) = \Omega(1/\varepsilon)$, so that this is the case we need to focus on. By Lemma 7.12, this is equivalently to the statement that $\mathcal{P}(\partial f^\star) > 0$. Since (as Hanneke, 2012, shows) the set V in CAL (in the realizable case) and RobustCAL (under Condition 2.3) has $\mathrm{DIS}(V)$ converging to $\partial f^\star$ (up to zero-probability differences) as the number of queries grows large, we see that after a sufficiently large number of label requests, a random point $x_1 \in \mathrm{DIS}(V)$ will be in $\partial f^\star$ with high probability; in fact, Hanneke [2012] shows $\mathcal{P}(\partial f^\star \setminus \partial_V f^\star) = 0$ almost surely, so that x_1 will also be in $\partial_V f^\star$ with high probability. If indeed $x_1 \in \partial_V f^\star$, we can actually *use* this fact to shrink the region in which the algorithm requests labels. Specifically, by definition of $\partial_V f^\star$, we know that there exist classifiers in V arbitrarily close to $f^\star$ which disagree on x_1. This has the interesting implication that, for any $y \in \mathcal{Y}$, defining $V[(x_1, y)] = \{h \in V : h(x_1) = y\}$, the set $V[(x_1, y)]$ has the property that, for almost every point x on which the classifiers in $V[(x_1, y)]$ *agree*, the agreed-upon classification will be $f^\star(x)$. Since this is true for both $y = -1$ and $y = +1$, we find that with conditional probability one (given V and x_1), if the next value of m in the algorithm has $X_m \notin \mathrm{DIS}(V[(x_1, -1)]) \cap \mathrm{DIS}(V[(x_1, +1)])$, then we can actually *infer* the value $f^\star(X_m)$. This is essentially the same property of $\mathrm{DIS}(V)$ that motivated the CAL and RobustCAL algorithms in Chapter 5, so that essentially the same reasoning can be applied to the methods that result from replacing $\mathrm{DIS}(V)$ in CAL and RobustCAL with the region $\mathrm{DIS}(V[(x_1, -1)]) \cap \mathrm{DIS}(V[(x_1, +1)])$, with one additional modification: namely, in the case that we have inferred the value $f^\star(X_m)$, we also use this value to eliminate from V any classifiers that disagree with the inferred classification; by the above reasoning about the correctness of these inferences, we may rest assured that this latter update will not remove $f^\star$ from V.

Note that the set $\mathrm{DIS}(V[(x_1, -1)]) \cap \mathrm{DIS}(V[(x_1, +1)])$ can sometimes be much smaller than $\mathrm{DIS}(V)$, so that these modified algorithms may request significantly fewer labels than the original CAL and RobustCAL algorithms. This is particularly interesting when $\mathcal{P}(\mathrm{DIS}(V[(x_1, -1)]) \cap \mathrm{DIS}(V[(x_1, +1)])) \to 0$ as the number of label requests grows large. Furthermore, when this is *not* the case, the above argument can be *iterated*. Specifically, when $\mathcal{P}(\mathrm{DIS}(V[(x_1, -1)]) \cap$

$\mathrm{DIS}(V[(x_1,+1)])) \nrightarrow 0$, the set $\mathrm{DIS}(V[(x_1,-1)]) \cap \mathrm{DIS}(V[(x_1,+1)])$ converges (almost surely) to the set $\partial_{\mathbb{C}[(x_1,-1)]} f^\star \cap \partial_{\mathbb{C}[(x_1,+1)]} f^\star$ (up to zero-probability differences), so that after a sufficiently large number of queries, a random point x_2 in $\mathrm{DIS}(V[(x_1,-1)]) \cap \mathrm{DIS}(V[(x_1,+1)])$ will be in $\partial_{\mathbb{C}[(x_1,-1)]} f^\star \cap \partial_{\mathbb{C}[(x_1,+1)]} f^\star$ with high probability; with slightly more thought, one can show it will also be in $\partial_{V[(x_1,-1)]} f^\star \cap \partial_{V[(x_1,+1)]} f^\star$ with high probability. If indeed $x_2 \in \partial_{V[(x_1,-1)]} f^\star \cap \partial_{V[(x_1,+1)]} f^\star$, we can again use this to reduce the region in which we request labels by noting this implies there are classifiers in $V[(x_1,y_1)][(x_2,y_2)] = \{h \in V : h(x_1) = y_1, h(x_2) = y_2\}$ arbitrarily close to $f^\star$, for every $y_1, y_2 \in \mathcal{Y}$. Thus, as above, for the next m obtained in the algorithm, with conditional probability one (given V, x_1, and x_2), if $X_m \notin \mathrm{DIS}(V[(x_1,y_1)][(x_2,y_2)])$ for some $y_1, y_2 \in \mathcal{Y}$, then the classification of X_m agreed-upon by the classifiers in $V[(x_1,y_1)][(x_2,y_2)]$ will be $f^\star(X_m)$; this is true of every $y_1, y_2 \in \mathcal{Y}$, so that we can infer $f^\star(X_m)$ when $X_m \notin \bigcap_{y_1,y_2 \in \mathcal{Y}} \mathrm{DIS}(V[(x_1,y_1)][(x_2,y_2)])$. Again, this is essentially the same property of $\mathrm{DIS}(V)$ used to motivate CAL and RobustCAL, so that it is natural to consider the active learning algorithms constructed by replacing the region $\mathrm{DIS}(V)$ in CAL and RobustCAL with the smaller region $\bigcap_{y_1,y_2 \in \mathcal{Y}} \mathrm{DIS}(V[(x_1,y_1)][(x_2,y_2)])$, with the additional modification that in the case that $X_m \notin \bigcap_{y_1,y_2 \in \mathcal{Y}} \mathrm{DIS}(V[(x_1,y_1)][(x_2,y_2)])$, we also eliminate from V any classifier that disagrees with the inferred value for $f^\star(X_m)$, confident that doing so will not remove $f^\star$ from V. This reasoning may be repeated as many times k as necessary to arrive at a partition of V into 2^k subsets with shrinking probability mass in the intersection of their regions of disagreement: i.e., $\mathcal{P}\left(\bigcap_{y_1,\ldots,y_k \in \mathcal{Y}} \mathrm{DIS}(V[(x_1,y_1)] \cdots [(x_k,y_k)])\right) \to 0$.

We can express the above argument more concisely in terms of *shattering*, since $\mathrm{DIS}(V)$ is merely the set of points x for which V shatters $\{x\}$, and given any $x_1 \in \mathrm{DIS}(V)$, the set $\mathrm{DIS}(V[(x_1,-1)]) \cap \mathrm{DIS}(V[(x_1,+1)])$ is merely the set of points x for which V shatters $\{x_1, x\}$, and so on. Thus, after k repetitions of the above argument, the region in which the algorithm would be requesting labels is simply the set of points x for which V shatters $\{x_1, \ldots, x_k, x\}$, where $\{x_1, \ldots, x_k\}$ is a (randomly constructed) collection of points shattered by V. Furthermore, the classification y the algorithm would infer for a point X_m for which V does not shatter $\{x_1, \ldots, x_k, X_m\}$ is the value $y \in \mathcal{Y}$ for which $V[(X_m, -y)]$ does not shatter $\{x_1, \ldots, x_k\}$, since the classification $(y_1, \ldots, y_k)$ of $(x_1, \ldots, x_k)$ that cannot be realized by classifiers in $V[(X_m, -y)]$ must have y as an agreed-upon classification of X_m by all classifiers in $V[(x_1, y_1)] \cdots [(x_k, y_k)] = \{h \in V : \forall i \leq k, h(x_i) = y_i\}$.

A few technical issues remain in the above description of this technique. First, as mentioned, we know that after sufficiently many label requests, a random point $x_1 \in \mathrm{DIS}(V)$ will be in $\partial_V f^\star$ with *high probability*, and more generally a random $\{x_1, \ldots, x_k\}$ shattered by V will be in $\lim_{\varepsilon \to 0} \{S \in \mathcal{X}^k : \mathrm{B}_V(f^\star, \varepsilon) \text{ shatters } S\}$ with high probability (as long as k is small enough for this latter set to have nonzero probability). Based on this, we concluded that the inferences of $f^\star(X_m)$ values would be accurate, with this same "high probability". However, to obtain label complexity guarantees with a favorable dependence on the confidence parameter δ, it is desirable for these "high" probabilities to be controllable; to achieve this effect, we can simply use the above argument repeatedly, sampling *many* random k-tuples $\{x_1, \ldots, x_k\}$ shattered by V, and then taking a *vote* among them on whether or not to request the label, and if not then which classification to infer. Since (after a sufficiently large number of queries), such a $\{x_1, \ldots, x_k\}$ will give the appropriate answer with probability greater than $1/2$, the vote will produce the appropriate decisions with a probability that can be made arbitrarily close to one, merely by taking a sufficiently large number of these random shatterable k-tuples. Generally, we can think of this as approximating the *probability* that a random shatterable k-tuple would vote in favor of requesting the label, and for simplicity we will simply express the algorithms below directly in terms of an unspecified estimator of these probabilities, with the understanding that in practice we could implement such estimators by this repeated voting process; the interested reader is referred to the original work of Hanneke [2012] for the details of these estimators.

The other detail we need to address before this technique can be implemented is the fact that it is difficult to detect which value of k will yield the required convergence of the probability of requesting a label.

To address this issue, we can simply try *every* value of k, using a fraction of the label budget for each such value; we start with the smaller values of k first, since the above argument indicates the inferred labels will be correct for smaller values as well, and this allows us to obtain that aforementioned "sufficiently large" number of label requests for the $f^{\star}(X_m)$ inferences to be correct by the time we reach the desirable larger value of k. Note that, in light of the above shatterability interpretation of this technique, we only need to try the values of k up to the point we are sampling shatterable d-tuples, since then (because no $d+1$ points are shattered by V) the algorithm will not request *any* further labels.

The formal details of the methods described above are provided below, for both the realizable case and noisy cases. We also describe the label complexity guarantees of this technique in terms of a generalization of the disagreement coefficient, naturally based on the rate of collapse of the probability $\mathrm{B}(f^{\star}, r)$ shatters a random k points, for an appropriate value of $k \in \mathbb{N}$. These results sometimes indicate strong improvements in label complexity compared to those of disagreement-based active learning methods.

8.5.1 The Realizable Case: The Shattering Algorithm

In the realizable case, we can apply the above reasoning to arrive at a modification of the CAL active learning algorithm, here referred to as the *Shattering* algorithm, originally due to Hanneke [2012]. As above, for any set $\mathcal{H}$ of classifiers, and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, denote $\mathcal{H}[(x, y)] = \{h \in \mathcal{H} : h(x) = y\}$. Additionally, for any $k \in \mathbb{N} \cup \{0\}$, denote by

$$\mathcal{S}^k(\mathcal{H}) = \{S \in \mathcal{X}^k : \mathcal{H} \text{ shatters } S\}.$$

Also, for $k \in \mathbb{N}$, denote by $\mathcal{P}^k$ the k-dimensional product measure (i.e., the joint distribution of $(X_1, \ldots, X_k)$), and also define $\mathcal{P}^0$ a probability measure on $\mathcal{X}^0 = \{()\}$ (which necessarily has $\mathcal{P}^0(\{()\}) = 1$ and $\mathcal{P}^0(\{\}) = 0$). The Shattering algorithm is then defined as follows.

Algorithm: **Shattering**(n)
0. $V \leftarrow \mathbb{C}$, $t \leftarrow 0$, $m \leftarrow 0$
1. For $k = 1, 2, \ldots, d+1$
2. While $t < (1 - 2^{-k})n$ and $m < 2^n$
3. $m \leftarrow m + 1$
4. If $\hat{\mathcal{P}}^{k-1}(S \in \mathcal{X}^{k-1} : S \cup \{X_m\} \in \mathcal{S}^k(V) | S \in \mathcal{S}^{k-1}(V)) \geq 1/2$
5. Request label Y_m and let $\hat{y} \leftarrow Y_m$, $t \leftarrow t+1$
6. Else $\hat{y} \leftarrow \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \hat{\mathcal{P}}^{k-1}\Big(\mathcal{X}^{k-1} \backslash \mathcal{S}^{k-1}(V[(X_m, -y)]) \Big| \mathcal{S}^{k-1}(V)\Big)$
7. $V \leftarrow V[(X_m, \hat{y})]$
8. Return any $\hat{h} \in V$

We assume ties will be broken in Step 6 in favor of a value for $\hat{y}$ with $V[(X_m, \hat{y})] \neq \emptyset$, to maintain the invariant that $V \neq \emptyset$. As mentioned above, the estimators $\hat{\mathcal{P}}^{k-1}$ can be implemented based on empirical averages of repeated samples of random k-tuples shattered by V; this only requires access to *unlabeled* data, and thus can achieve arbitrarily good precision with high confidence, without affecting the label complexity. The interested reader is referred to the original discussion of Hanneke [2012] for the details of such estimators. As was the case for CAL, in practice one would maintain the set V *implicitly* as a set of constraints, and the references to V in Steps 4, 6, and 8 can then be expressed as constraint satisfaction problems. For brevity, we leave the details of this alternative description of the Shattering algorithm as an exercise for the reader.

To quantify the label complexity of the Shattering algorithm, Hanneke [2012] introduces the following natural generalizations of the disagreement core and disagreement coefficient.

Definition 8.9. For any classifier h and any $k \in \mathbb{N} \cup \{0\}$, let

$$\partial^k h = \lim_{\varepsilon \to 0} \mathcal{S}^k(\mathrm{B}(h, \varepsilon)),$$

and for $r_0 \geq 0$, define

$$\theta_h^{(k)}(r_0) = \sup_{r > r_0} \frac{\mathcal{P}^k\left(\mathcal{S}^k\left(\mathrm{B}(h, r)\right)\right)}{r} \vee 1.$$

Also denote by $\tilde{d}_h = \min\left\{k \in \mathbb{N} : \mathcal{P}^k\left(\partial^k h\right) = 0\right\}$, and define

$$\tilde{\theta}_h(r_0) = \theta_h^{(\tilde{d}_h)}(r_0).$$

When $h = f^\star$, abbreviate these as $\theta^{(k)}(r_0) = \theta_{f^\star}^{(k)}(r_0)$, $\tilde{d} = \tilde{d}_{f^\star}$, and $\tilde{\theta}(r_0) = \tilde{\theta}_{f^\star}(r_0)$.

The set $\partial^k h$ is referred to as the k-dimensional shatter core of h with respect to $\mathbb{C}$ under $\mathcal{P}$, and the quantity $\theta_h^k(\cdot)$ is called the order-k disagreement coefficient of h with respect to $\mathbb{C}$ under $\mathcal{P}$. Note that $\tilde{d}_h \leq d+1$, so that $\tilde{d}_h$ is always well-defined and finite when the VC dimension of $\mathbb{C}$ is finite. Also note that $\tilde{\theta}(\cdot) \leq \theta(\cdot)$; indeed, $\theta(\cdot) = \theta^{(1)}(\cdot)$. Additionally, as in Lemma 7.12, we have $\theta_h^{(k)}(\varepsilon) = o(1/\varepsilon)$ iff $\mathcal{P}^k(\partial^k h) = 0$. Therefore, unlike $\theta(\varepsilon)$, we *always* have $\tilde{\theta}(\varepsilon) = o(1/\varepsilon)$ due to our choice of $\tilde{d}$.

Hanneke [2012] bounds $\tilde{\theta}(\varepsilon)$ for several different learning problems. For instance, for $\mathbb{C}$ the class of linear separators (Example 3) and $\mathcal{P}$ a uniform distribution on a sphere, Hanneke [2012] shows $\tilde{\theta}_h(\varepsilon) = O(1)$ for every $h \in \mathbb{C}$; recall that this is not the case for $\theta_h(\varepsilon)$, particularly when the separating hyperplane corresponding to h does not intersect the sphere (in which case $\theta_h(\varepsilon) = 1/\varepsilon$).

The label complexity of the Shattering algorithm can be bounded in terms of $\tilde{\theta}$, as stated in the following theorem of Hanneke [2012] (the version presented here has slightly sharper logarithmic factors, due to using Lemma 3.1 in place of a weaker bound used in the original proof). We omit the formal details of the proof for brevity, referring the interested reader to the original article of Hanneke [2012].

Theorem 8.10. Shattering achieves a label complexity Λ such that, for $\mathcal{P}_{XY}$ in the realizable case, and a constant $c \in (1, \infty)$, $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \leq c\tilde{\theta}(\varepsilon)\left(d\mathrm{Log}(\theta(\varepsilon)) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right)\mathrm{Log}(1/\varepsilon).$$

Aside from constant factors, this is never worse than Theorem 5.1, and since $\tilde{\theta}(\varepsilon)$ is *always* $o(1/\varepsilon)$, it is often significantly better. As was the case in Theorem 5.1, the logarithmic factors can often be improved. Note that the constant c in Theorem 8.10 may depend on $\mathbb{C}$ and $\mathcal{P}_{XY}$ (see Hanneke, 2012, for an explicit description of this dependence).

Unlike the disagreement coefficient analysis of CAL, the bound in Theorem 8.10 does not always provide a tight characterization of the label complexity of the Shattering algorithm, even up to constants and logarithmic factors. Much of this issue comes from the fact that the label complexity bound in Theorem 8.10 is essentially proportional to a bound on the number of labels that would need to be requested while $k = \tilde{d}$ in the Shattering algorithm before reaching some particular value of m. This is valid, since (by the reasoning above) the labels inferred in Step 6 while $k \leq \tilde{d}$ are correct (for sufficiently large n); however, in some cases, the labels inferred in Step 6 may even be correct for some values of k *larger* than $\tilde{d}$, so that we could instead bound the (smaller) number of labels the algorithm would request under this larger k before reaching that value of m. For instance, this is the case when $\mathbb{C}$ is the class of linear separators (Example 3) in at least 2 dimensions, $\mathcal{P}$ is a uniform distribution inside a ball of nonzero radius, and $f^{\star}$ is a linear separator with separating hyperplane not intersecting this ball; in this case, as discussed in Chapter 7, $1 \ll \theta(\varepsilon) \ll 1/\varepsilon$, so that $\tilde{d} = 1$ (by Lemma 7.12), and thus $\tilde{\theta}(\varepsilon) = \omega(1)$, but some quick reasoning reveals that after a sufficiently large number of label requests while $k = 1$, $\mathrm{DIS}(V)$ will be structured in such a way that the inferences in Step 6 will be correct for $k = 2$, and in fact that the algorithm will not request any labels while $k \geq 2$. At present, there is no known concise general characterization of the label complexity of the Shattering algorithm that takes these more subtle issues into account.

8.5.2 The Noisy Case

The same ideas leading to the Shattering algorithm can also be applied in the presence of classification noise. Specifically, we can make the Shattering algorithm robust to noise, using the same approach taken in Chapter 5 to arrive at RobustCAL as a noise-robust variant of CAL. Formally, we can state this algorithm as follows, here referred to as RobustShattering (where, as in RobustCAL, $\delta_m = \delta/(\log_2(2m))^2$).

Algorithm: $\mathbf{RobustShattering}_{\delta}(n)$
0. $m \leftarrow 0$, $Q \leftarrow \{\}$, $V \leftarrow \mathbb{C}$
1. For $k = 1, 2, \ldots, d+1$
2. While $|Q| < (1 - 2^{-k})n$ and $m < 2^n$
3. $m \leftarrow m + 1$
4. If $\hat{\mathcal{P}}^{k-1}(S \in \mathcal{X}^{k-1} : S \cup \{X_m\} \in \mathcal{S}^k(V) | S \in \mathcal{S}^{k-1}(V)) \geq 1/2$
5. Request the label Y_m; let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$
6. Else $\hat{y} \leftarrow \operatorname*{argmax}_{y \in \mathcal{Y}} \hat{\mathcal{P}}^{k-1}\Big(\mathcal{X}^{k-1} \setminus \mathcal{S}^{k-1}(V[(X_m, -y)])\Big| \mathcal{S}^{k-1}(V)\Big)$
7. $V \leftarrow V[(X_m, \hat{y})]$
8. If $\log_2(m) \in \mathbb{N}$
9. $V \leftarrow \left\{h \in V : \left(\mathrm{er}_Q(h) - \min_{g \in V} \mathrm{er}_Q(g)\right) |Q| \leq U(m, \delta_m) m\right\}$
10. Return any $\hat{h} \in V$

As in the Shattering algorithm, we assume ties will be broken in Step 6 in favor of a value for $\hat{y}$ with $V[(X_m, \hat{y})] \neq \emptyset$, to maintain the invariant that $V \neq \emptyset$. The above algorithm comes from the work of Hanneke [2012], though the original formulation of the algorithm uses a data-dependent estimator $\hat{U}$ in place of U in Step 9, so that the algorithm has no direct dependence on $\mathcal{P}_{XY}$ aside from access to the data. This is the same data-dependent estimator alluded to above for RobustCAL; again, we omit the details of $\hat{U}$ here for brevity, referring the interested reader to the original presentation of Hanneke [2012]. Also, as was the case for the Shattering algorithm, the estimators $\hat{\mathcal{P}}$ referenced in RobustShattering can be implemented based purely on unlabeled samples [see Hanneke, 2012]. As in RobustCAL, in practice, one would typically maintain the set V implicitly as a set of constraints, so that the various steps involving V in the algorithm would be implemented by solving constraint satisfaction or constrained optimization problems.

The following theorem, regarding the label complexity of RobustShattering, is due to Hanneke [2012] (though with slightly sharper logarithmic factors here, due to using Lemma 3.1 in place of a weaker bound in the original proof). We omit the proof here for brevity.

Theorem 8.11. For any $\mathcal{P}_{XY}$, for a and α as in Condition 2.3, for a $\mathcal{P}_{XY}$-dependent constant $c \in (1, \infty)$, for any $\delta \in (0, 1)$, RobustShattering$_\delta$ achieves a label complexity Λ such that, $\forall \varepsilon \in (0, 1)$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \\ \leq ca^2\tilde{\theta}\left(a\varepsilon^\alpha\right)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\mathrm{Log}\left(\theta\left(a\varepsilon^\alpha\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

As with Theorem 8.10, this label complexity is never worse (aside from constant factors) than that of RobustCAL in Theorem 5.4, and can often be significantly better when $\tilde{d} > 1$. As in Theorems 5.4 and 8.10, the logarithmic factors can be refined in many cases.

It is also worth noting that the shattering-based active learning strategy is compatible with the discussion of surrogate losses in Chapter 6. Specifically, one can formulate a variant of RobustShattering that relaxes the optimizations and constraints involving the 0-1 loss, substituting an arbitrary classification-calibrated surrogate loss. With a few additional modifications to the algorithm, one can obtain bounds on the label complexity, generalizing Theorem 8.11, in much the same way we generalized Theorem 5.4 to arrive at Theorem 6.8.

8.6 Active Learning Always Helps

We have already seen a number of results on the label complexity of active learning methods, for various specific types of hypothesis classes and distributions, which cannot hold for any passive learning methods. However, there is a question of how general of a claim we can make regarding the ability of active learning methods to provide improvements over passive learning methods. One strong type of guarantee we might hope for is that, if a given passive learning method has label complexity Λ_p, then the active learning algorithm will have label complexity $o(\Lambda_p(\varepsilon, \delta, \mathcal{P}_{XY}))$, for all $\mathcal{P}_{XY}$ of a given type (e.g., in the realizable case). This subsection discusses some quite general results of this type, based on the works of Balcan, Hanneke, and Vaughan [2010] and Hanneke [2009b, 2012].

To simplify the discussion, rather than comparing label complexity functions as defined in Section 2.2, we will instead focus on the label complexity of achieving *expected* error rate ε. Specifically, we say an active learning algorithm $\mathcal{A}_a$ achieves an expected-error label complexity $\Lambda_a(\cdot,\cdot)$ if, for every $\varepsilon \in [0,1]$, every distribution $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, and every integer $n \geq \Lambda_a(\varepsilon, \mathcal{P}_{XY})$, if $\hat{h}$ is the classifier produced by running $\mathcal{A}_a$ with budget n, then $\mathbb{E}[\mathrm{er}(\hat{h})] \leq \varepsilon$. Furthermore, as before, we extend this definition to passive learning algorithms $\mathcal{A}_p$ by applying this definition to the simple active learning algorithm that requests the first n labels $Y_1, \ldots, Y_n$ and returns $\mathcal{A}_p(\mathcal{Z}_n)$, where n is the given label budget. The reason we focus on the expected-error label complexity, rather than the type that includes a confidence parameter δ, is that it simplifies the discussion of asymptotic analysis to have only a single variable (i.e., ε), rather than a function of two variables. The results we discuss here also hold for the original $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$ functions, if we consider δ to be held constant when we study the behavior of the label complexity as $\varepsilon \to 0$ (in fact, they also become much easier to establish in this case).

For a given expected-error label complexity Λ_p achieved by a given passive learning algorithm, we will be interested in determining whether there is an active learning algorithm achieving an expected-error label complexity Λ_a that is generally much smaller than Λ_p. However, before making this formal, we first need to restrict the types of distributions $\mathcal{P}_{XY}$ we require this statement to hold for. For instance, if there is some $x \in \mathcal{X}$ with $\mathcal{P}(\{x\}) = 1$, we cannot possibly hope to always have $\Lambda_a(\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$, since the behavior of any active learning algorithm will be distributionally equivalent to some passive learning algorithm in this case. To resolve this issue, we need to have a notion of a *nontrivial* distribution $\mathcal{P}_{XY}$. For this purpose, we define the set $\mathrm{Nontrivial}(\Lambda_p)$ as the set of all distribution $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that, $\forall k \in \mathbb{N}$, $\Lambda_p(\varepsilon + \inf_h \mathrm{er}(h), \mathcal{P}_{XY}) = \omega((\mathrm{Log}(1/\varepsilon))^k)$; this definition is reasonable, since $\mathrm{polylog}(1/\varepsilon)$ label complexities are usually thought of as being quite small already (though Hanneke, 2012, also explores weaker notions of nontriviality). Additionally, for now we just focus on $\mathcal{P}_{XY}$ in the realizable case, and discuss noisy cases below. The following theorem was proven by Hanneke [2012].

Theorem 8.12. If $d < \infty$, for any expected-error label complexity Λ_p achieved by a passive learning algorithm, there exists an active learning algorithm achieving an expected-error label complexity Λ_a such that, for all $\mathcal{P}_{XY} \in \text{Nontrivial}(\Lambda_p)$ in the realizable case, $\forall c > 1$, $\Lambda_a(c\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$.

This theorem essentially says that for any passive learning algorithm, there is an active learning algorithm that is asymptotically much better. The theorem does admit a slight loss in the ε argument to Λ_a by a factor of c; this is typically not significant (particularly if $\Lambda_p(\varepsilon, \mathcal{P}_{XY}) = \text{poly}(1/\varepsilon)$, as is typically the case), but nonetheless it is presently not known whether there is a method achieving this result with $c = 1$.

We have already seen two different techniques that can be used to approach guarantees of this type: namely, Theorem 7.31 and the Shattering algorithm. The first approach (due to Balcan, Hanneke, and Vaughan, 2010) is via Theorem 7.31, which states that one can decompose $\mathbb{C}$ into disjoint subsets, each of which has $o(1/\varepsilon)$ disagreement coefficients (with respect to a countable dense subset of that subclass). By running a variant of CAL on each of these subsets (with an appropriate label budget for each), using the resulting classifier to classify the unlabeled data points processed in each case, feeding (portions of) these labeled data sets into the passive learning algorithm $\mathcal{A}_p$ achieving Λ_p, and choosing among the resulting classifiers with a kind of model selection procedure (similar to the ActiveSelect subroutine described below), one is able to achieve an expected-error label complexity $\Lambda_a(\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon/c, \mathcal{P}_{XY}))$ for any constant $c > 1$, for $\mathcal{P}_{XY} \in \text{Nontrivial}(\Lambda_p)$ in the realizable case [Balcan, Hanneke, and Vaughan, 2010]. However, the decomposition from Theorem 7.31 used by this algorithm is $\mathcal{P}$-dependent, and there is no obvious way to supplant this dependence with data-dependent estimators, so that we would require direct access to the distribution $\mathcal{P}$ in order to run the algorithm.

The second approach (due to Hanneke, 2009b, 2012) uses a variant of the Shattering algorithm, and has no direct dependence on $\mathcal{P}$. Specifically, suppose $\mathcal{A}_p$ is the passive learning algorithm achieving expected-error label complexity Λ_p, and consider the following algorithm (the function $s(n)$ appearing in it is discussed below).

Algorithm: **ShatteringActivizer**($\mathcal{A}_p$, n)
0. $V \leftarrow \mathbb{C}$, $t \leftarrow 0$, $m \leftarrow 0$
1. For $k = 1, 2, \ldots, d+1$
2. Let $\mathcal{L}_k \leftarrow \{\}$
3. While $t < (2 - 2^{-k} - 2^{1-k})n/4$ and $m < 2^n$
4. $m \leftarrow m + 1$
5. If $\hat{\mathcal{P}}^{k-1}(S \in \mathcal{X}^{k-1} : S \cup \{X_m\} \in \mathcal{S}^k(V) | S \in \mathcal{S}^{k-1}(V)) \geq 1/2$
6. Request the label Y_m and let $\hat{y} \leftarrow Y_m$, $t \leftarrow t + 1$
7. Else $\hat{y} \leftarrow \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \hat{\mathcal{P}}^{k-1}\Big(\mathcal{X}^{k-1} \setminus \mathcal{S}^{k-1}(V[(X_m, -y)])\Big|\mathcal{S}^{k-1}(V)\Big)$
8. $V \leftarrow V[(X_m, \hat{y})]$
9. $\hat{\Delta}^{(k)} \leftarrow \hat{\mathcal{P}}(x : \hat{\mathcal{P}}^{k-1}(S \in \mathcal{X}^{k-1} : S \cup \{x\} \in \mathcal{S}^k(V) | S \in \mathcal{S}^{k-1}(V)) \geq 1/2)$
10. Do $\lfloor 2^{-k} n / (5\hat{\Delta}^{(k)}) \rfloor$ times
11. $m \leftarrow m + 1$
12. If $\hat{\mathcal{P}}^{k-1}(S \in \mathcal{X}^{k-1} : S \cup \{X_m\} \in \mathcal{S}^k(V) | S \in \mathcal{S}^{k-1}(V)) \geq 1/2$ and $t < (1 - 2^{-k})n/2$
13. Request the label Y_m and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t + 1$
14. Else $\hat{y} \leftarrow \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \hat{\mathcal{P}}^{k-1}\Big(\mathcal{X}^{k-1} \setminus \mathcal{S}^{k-1}(V[(X_m, -y)])\Big|\mathcal{S}^{k-1}(V)\Big)$
15. $\mathcal{L}_k \leftarrow \mathcal{L}_k \cup \{(X_m, \hat{y})\}$, $V \leftarrow V[(X_m, \hat{y})]$
16. Return ActiveSelect($\{\mathcal{A}_p(\mathcal{L}_1), \ldots, \mathcal{A}_p(\mathcal{L}_{d+1})\}$, $\lfloor n/2 \rfloor$, m)

Subroutine: **ActiveSelect**($\{h_1, \ldots, h_N\}$, n, m)
0. For each $j, k \in \{1, \ldots, N\}$ s.t. $j < k$,
1. Let R_{jk} be the set of the first $\left\lfloor \frac{n}{j(N-j)\ln(eN)} \right\rfloor$ indices $m' > m$ with $h_j(X_{m'}) \neq h_k(X_{m'})$ (if such indices exist)
2. Request the labels $Y_{m'}$ for each $m' \in R_{jk}$, and let $Q_{jk} = \{(X_{m'}, Y_{m'}) : m' \in R_{jk}\}$
3. Return $h_{\hat{k}}$, where $\hat{k} = \max\{k \in \{1, \ldots, N\} : \max_{j<k} \operatorname{er}_{Q_{jk}}(h_k) \leq (1 + s(n))(1/2)\}$

This algorithm can be used to prove Theorem 8.12 by the following reasoning (for a complete formal proof, see the original work of Hanneke, 2012). As discussed in Section 8.5, for some value of k (in particular, for $k = \tilde{d}$), for any sufficiently large value of n, the probability of requesting a label Y_m in Step 6 or 13 will be shrinking to 0 (with high probability) as m grows large, and the labels $\hat{y}$ inferred in Steps 7 and 14 will have $\hat{y} = f^\star(X_m)$ (again, with high probability). For an appropriate choice of the estimator $\hat{\mathcal{P}}$ in Step 9, this implies $\hat{\Delta}^{(k)}$ will be shrinking to 0 as n grows large, so that $|\mathcal{L}_k| = \omega(n)$ in Step 16, while $\mathrm{er}_{\mathcal{L}_k}(f^\star) = 0$. In other words, $\mathcal{L}_k$ is a correctly-labeled data set. Also note that, since $|\mathcal{L}_k|$ is set in Step 10 before processing any of the data points X_m that compose $\mathcal{L}_k$ itself, the sequence of unlabeled samples X_m used to construct $\mathcal{L}_k$ in Step 15 are conditionally i.i.d. (each with distribution $\mathcal{P}$) given $|\mathcal{L}_k|$. Thus, aside from the effect of the constraint on t in Step 12, $\mathcal{L}_k$ is conditionally i.i.d. (each entry with distribution $\mathcal{P}_{XY}$) given $|\mathcal{L}_k|$, so that having $|\mathcal{L}_k| \geq \Lambda_p(\varepsilon, \mathcal{P}_{XY})$ would imply $\mathbb{E}[\mathrm{er}(\mathcal{A}_p(\mathcal{L}_k))||\mathcal{L}_k|] \leq \varepsilon$. A simple Chernoff bound argument implies that, with high probability, the constraint on t in Step 12 will never be violated. Thus, since $|\mathcal{L}_k| = \omega(n)$, if $\mathcal{P}_{XY} \in \mathrm{Nontrivial}(\Lambda_p)$ and $\forall \varepsilon > 0$, $\Lambda_p(\varepsilon, \mathcal{P}_{XY}) < \infty$, in the realizable case, an algorithm that returns $\mathcal{A}_p(\mathcal{L}_k)$ would have expected-error label complexity Λ_a with $\Lambda_a(c\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$ for any $c > 1$ (where the factor of c is needed due to the "high probability" qualification on each of the above claims; it turns out these probabilities are each $1 - o(\varepsilon)$ when $n = \omega(\mathrm{Log}(1/\varepsilon))$).

Since it is not always possible to determine for which k this argument holds from data alone (without using prohibitively many additional label requests), we simply calculate $\mathcal{A}_p(\mathcal{L}_k)$ for *every* value of k, and then apply the ActiveSelect procedure to select from among the resulting classifiers. ActiveSelect runs a kind of tournament, requesting a number of labels for points in the regions of disagreement of pairs of these classifiers. Supposing $s(n) \in (0, 1/4]$, and $s(n) = O(n^{-\tau})$ for some $\tau \in [0, 1/2)$, By a Chernoff bound, with probability $1 - \exp\{-\Omega(s(n)^2 n)\}$, if h_{k^*} has the lowest error rate among these classifiers, then for every $j < k^*$, $\mathrm{er}_{Q_{jk^*}}(h_{k^*}) \leq (1 + s(n))(1/2)$, while any $k' > k^*$ with $\mathrm{er}(h_{k'}) > (1 + 4s(n))\mathrm{er}(h_{k^*})$ *will* have $\mathrm{er}_{Q_{k^*k'}}(h_{k'}) > (1 + s(n))(1/2)$. Therefore, with high probability, the classifier returned by ActiveSelect has error rate not much larger than $\min_k \mathrm{er}(h_k)$. Altogether, we see that ShatteringActivizer$(\mathcal{A}_p, \cdot)$ achieves an expected-error label complexity Λ_a such that $\Lambda_a(c\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$ for any $c > 1$, for any $\mathcal{P}_{XY} \in \text{Nontrivial}(\Lambda_p)$ in the realizable case having $\forall \varepsilon > 0, \Lambda_p(\varepsilon, \mathcal{P}_{XY}) < \infty$.

It is straightforward to extend this to have $\Lambda_a(c\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$ for those $\mathcal{P}_{XY}$ in the realizable case with $\Lambda_p(\varepsilon, \mathcal{P}_{XY}) = \infty$ for some values $\varepsilon > 0$. Specifically, we can simply let $\hat{h}_1 = \text{ShatteringActivizer}(\mathcal{A}_p, \lfloor n/3 \rfloor)$, let $\hat{h}_2 = \text{ERM}(\mathbb{C}, \mathcal{Z}_{\lfloor n/3 \rfloor})$ (after requesting labels $Y_1, \ldots, Y_{\lfloor n/3 \rfloor}$), and then return the classifier $\hat{h} = \text{ActiveSelect}(\{\hat{h}_1, \hat{h}_2\}, \lceil n/3 \rceil)$. The resulting method loses only a constant factor in $\Lambda_a(\varepsilon, \mathcal{P}_{XY})$ for those $\mathcal{P}_{XY}$ with finite $\Lambda_p(\varepsilon, \mathcal{P}_{XY})$ values, while it guarantees $\Lambda_a(\varepsilon, \mathcal{P}_{XY}) < \infty$ ($\forall \varepsilon > 0$) for all $\mathcal{P}_{XY}$ in the realizable case, including those with $\Lambda_p(\varepsilon, \mathcal{P}_{XY}) = \infty$ for some values $\varepsilon > 0$.

Since claims of the type in Theorem 8.12, and the methods achieving them, can be interesting to study in a variety of contexts and under a variety of conditions, Hanneke [2012] abstracts the study of this type of behavior into a general reduction-style framework. Specifically, define the notion of an *active meta-algorithm* $\mathcal{A}_a(\cdot, \cdot)$ as taking two arguments, a passive algorithm and a label budget, with the property that for any passive algorithm $\mathcal{A}_p$, $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ is an active learning algorithm; then we say an active meta-algorithm $\mathcal{A}_a$ *activizes* a passive algorithm $\mathcal{A}_p$ for $\mathbb{C}$ if the active learning algorithm $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ achieves an expected-error label complexity Λ_a such that, for any expected-error label complexity Λ_p achieved by $\mathcal{A}_p$, for all $\mathcal{P}_{XY} \in \text{Nontrivial}(\Lambda_p)$ in the realizable case with $\forall \varepsilon > 0, \Lambda_p(\varepsilon, \mathcal{P}_{XY}) < \infty$, $\exists c \in [1, \infty)$ such that $\Lambda_a(c\varepsilon, \mathcal{P}_{XY}) = o(\Lambda_p(\varepsilon, \mathcal{P}_{XY}))$. An active meta-algorithm $\mathcal{A}_a$ is called a *universal activizer for* $\mathbb{C}$ if it activizes *every* passive learning algorithm $\mathcal{A}_p$ for $\mathbb{C}$. Thus, the above reasoning indicates that ShatteringActivizer is a universal activizer for $\mathbb{C}$ (if $d < \infty$).

At present, there are many interesting open problems regarding the conditions under which universal activizers for $\mathbb{C}$ exist. For instance, Hanneke [2012] shows that certain classes with $d = \infty$ still have universal activizers for them (e.g., if $\mathbb{C} = \bigcup_i \mathbb{C}_i$, where each i has $\text{vc}(\mathbb{C}_i) < \infty$). However, it is not known whether there is a universal activizer for the class of *all* classifiers (referred to simply as a *universal activizer*, since it is universal for *every* hypothesis class).

There are also important questions on the existence of activizers when we remove the restriction of $\mathcal{P}_{XY}$ being in the realizable case. Hanneke [2012] proves that we typically should not expect *universal* activizers to exist in the general noisy case. However, it might still be the case that there are activizers for some broad family of *reasonable* passive learning methods; in particular, Hanneke [2012] conjectures that, when $d < \infty$, there is an activizer for some $\mathrm{ERM}(\mathbb{C}, \cdot)$, even without any restrictions on $\mathcal{P}_{XY}$ (i.e., the so-called *agnostic* case).

8.7 Verifiability

The issue of *verifiability* of low error rate is an important and nontrivial one for active learning. In passive learning, for any classifier h with $\mathrm{er}(h) \leq \varepsilon/2$, we can easily verify at least that $\mathrm{er}(h) \leq \varepsilon$ simply by taking $\lesssim (1/\varepsilon)\mathrm{Log}(1/\delta)$ random labeled samples $\mathcal{L}$ and checking whether $\mathrm{er}_{\mathcal{L}}(h) \leq (2/3)\varepsilon$. Thus, since the number of samples required for passive learning is typically $\Omega(1/\varepsilon)$ anyway, we see that the number of samples required for both learning (to error rate ε with probability $1-\delta$) *and* verification of success is not much larger than the label complexity of passive learning (i.e., without verification). However, this convenience is not available to us in active learning, since $(1/\varepsilon)\mathrm{Log}(1/\delta)$ samples would be considered a relatively large number, compared to the number of labels needed for learning (which, as we have seen, is often $o(1/\varepsilon)$). Furthermore, it turns out the number of labels needed for verification can sometimes be substantially *larger* than the number of labels sufficient for learning (without verification). To formalize these observations, consider the following definition from Balcan, Hanneke, and Vaughan [2010].

Definition 8.13. An active learning algorithm $\mathcal{A}$ achieves a verifiable label complexity Λ for the realizable case if there exists a value $\hat{\varepsilon}_{n,\delta}$ for each $n \in \mathbb{N}$ and $\delta \in (0,1)$, where the value of each $\hat{\varepsilon}_{n,\delta}$ is determined only by $\mathcal{Z}_{\mathbf{X}}$ and the labels Y_t requested during the execution of $\mathcal{A}(n)$, such that, for any $\varepsilon, \delta \in (0,1)$, any distribution $\mathcal{P}_{XY}$ in the realizable case, and any $n \in \mathbb{N}$, with probability at least $1-\delta$, the classifier $\hat{h}_n = \mathcal{A}(n)$ satisfies $\mathrm{er}(\hat{h}_n) \leq \hat{\varepsilon}_{n,\delta}$, and if $n \geq \Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$, then $\hat{\varepsilon}_{n,\delta} \leq \varepsilon$

as well.

The requirement that a data-dependent error estimate $\hat{\varepsilon}_{n,\delta}$ exists places a significant restriction on the verifiable label complexities that can be achieved. For instance, consider the problem of learning interval classifiers (Example 2), with $\mathcal{P}$ a uniform distribution over $[0,1]$, and $f^\star = \mathbb{1}^{\pm}_{[a,a]}$ for some $a \in (0,1)$. In this case, Balcan, Hanneke, and Vaughan [2010] prove that any verifiable label complexity Λ for the realizable case has $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) = \Omega(1/\varepsilon)$. The reason is that, if $\mathbb{P}(\mathrm{er}(\hat{h}_n) \leq \hat{\varepsilon}_{n,\delta} \leq \varepsilon) \geq 1 - \delta$ in this case, for some $n < c/\varepsilon$ (for an appropriate $c \in (0,1)$), then there would be greater than δ probability that, for the problem of learning with some alternative target function (specifically, an interval of width 3ε), $\mathcal{A}(n)$ would produce the same classifier $\hat{h}_n$ and error estimate value $\hat{\varepsilon}_{n,\delta}$ (since none of the requested labels would be positive), which in this case would have $\mathrm{er}(\hat{h}_n) > \varepsilon \geq \hat{\varepsilon}_{n,\delta}$; thus, any such $\hat{\varepsilon}_{n,\delta}$ would not satisfy the requirement of Definition 8.13. In particular, since there is a passive learning algorithm achieving verifiable label complexity $\lesssim (1/\varepsilon)\mathrm{Log}(1/\delta)$ in the realizable case for this problem [Haussler, Littlestone, and Warmuth, 1994], we immediately see that general results claiming strong advantages of active learning over passive learning in the realizable case are *not* possible for the verifiable label complexity, unlike the (unverifiable) label complexity from Definition 2.1 [Balcan, Hanneke, and Vaughan, 2010, Hanneke, 2012].

Many of the techniques described in this article are equally-effective at bounding the verifiable label complexity. For instance, note that the bounds on the label complexity Λ achieved by CAL given in Theorem 5.1 hold for $\hat{h}_n$ an arbitrary element of the final version space V; thus, if we take $\hat{\varepsilon}_{n,\delta} = \sup_{h,g \in V} \mathcal{P}(x : h(x) \neq g(x))$ (or rather, a good estimate thereof based on unlabeled data) at the conclusion of the algorithm, we see that the verifiable label complexity can be bounded by $\Lambda(\varepsilon/2, \delta, \mathcal{P}_{XY})$, so that Theorem 5.1 also holds for the verifiable label complexity (after appropriate modifications to the constant factors in the bound). A similar argument applies to the Splitting algorithm, so that Theorem 8.1 also holds for the verifiable label complexity. For the ActiveHalving algorithm, Theorem 8.8 remains valid for the verifiable label complexity, simply taking $\hat{\varepsilon}_{n,\delta} = (1/m)\mathrm{Log}(4nm/\delta)$ if the minimizing value of the summation in Step 8 is 0, and otherwise taking $\hat{\varepsilon}_{n,\delta} = 1$.

However, this does *not* apply to the label complexity analysis of the Shattering algorithm described in Section 8.5. In particular, the final version space V in that algorithm might no longer contain $f^\star$, so that $\sup_{h,g\in V} \mathcal{P}(x : h(x) \neq g(x))$ does not necessarily upper-bound $\mathrm{er}(\hat{h})$. Indeed, the label complexity bound in Theorem 8.10 typically does not bound the verifiable label complexity of that algorithm. For instance, this must be the case for $\mathbb{C}$ the class of interval classifiers, $\mathcal{P}$ the uniform distribution over $[0,1]$, and $f^\star \in \mathbb{C}$ with $\mathcal{P}(x : f^\star(x) = +1) = 0$; $\tilde{\theta}(\varepsilon) = O(1)$ for this problem, so that the label complexity bound in Theorem 8.10 has a dependence on ε of $O((\mathrm{Log}(1/\varepsilon))^2)$, and is therefore smaller than the aforementioned $\Omega(1/\varepsilon)$ lower bound on the verifiable label complexity for this scenario.

The verifiable label complexity can also be formalized for noisy settings, simply by requiring that, for any distribution $\mathcal{P}_{XY}$, for any $n \in \mathbb{N}$, with probability at least $1-\delta$, $\mathrm{er}(\hat{h}_n) - \nu \leq \hat{\varepsilon}_{n,\delta}$, and if $n \geq \Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY})$ then $\hat{\varepsilon}_{n,\delta} \leq \varepsilon$ as well. Once again, some of the results discussed in this article hold for this notion of verifiable label complexity as well. For instance, in RobustCAL, we can let $m' = 2^{\lfloor \log_2(m) \rfloor}$ for the final value of m in the algorithm, and take $\hat{\varepsilon}_{n,\delta} = 2U(m', \delta_{m'})$ (or rather, the data-dependent estimator of this, $2\hat{U}(m', \delta_{m'})$, alluded to in Section 5.2); the proof of Theorem 5.4 already establishes that this has the required properties, so that Theorem 5.4 remains valid for the verifiable label complexity as well.

8.7.1 Self-Verifying Active Learning

The notion of the verifiable label complexity is closely related to the idea of a *self-verifying* (or *self-terminating*) active learning algorithm. Specifically, consider a type of algorithm $\mathcal{A}(\cdot,\cdot)$ which takes two arguments, $\varepsilon, \delta \in (0,1)$, requests any number of labels (where the number can vary adaptively, based on the observed data and label responses), and then returns a classifier $\hat{h}$, with the property that, $\forall \varepsilon, \delta \in (0,1)$, with probability at least $1-\delta$, the classifier $\hat{h} = \mathcal{A}(\varepsilon, \delta)$ satisfies

$\mathrm{er}(\hat{h}) \leq \varepsilon$ (or $\mathrm{er}(\hat{h}) - \mathrm{er}(f^\star) \leq \varepsilon$, in the nonrealizable case). A significant number of active learning algorithms in the literature are expressed as this type of self-verifying algorithm, including the original version of the Splitting algorithm [Dasgupta, 2005], the original version of the ActiveHalving algorithm and its noise-robust counterpart [Hanneke, 2007a], and the original A^2 algorithm mentioned in Section 5.3 [Balcan, Beygelzimer, and Langford, 2006, 2009].

One can convert either type of algorithm (budget-based or self-verifying) into the other, typically with the number of labels requested by the self-verifying algorithm of roughly the same magnitude as the verifiable label complexity of the corresponding budget-based algorithm [Balcan, Hanneke, and Vaughan, 2010]. Given any self-verifying active learning algorithm $\mathcal{A}$, and given a budget n, we can simply run $\mathcal{A}(2^{-i}, \delta/(i+1)^2)$ for increasing values of $i \in \mathbb{N}$ until it requests a total of n labels, and then return the classifier produced by the last execution of the algorithm that ran to completion, and take the corresponding 2^{-i} value as the value of $\hat{\varepsilon}_{n,\delta}$. Since we expect the number of labels requested by $\mathcal{A}(2^{-i}, \delta/(i+1)^2)$ to be increasing in i, and since this number typically has only a polylog dependence on the second argument, we expect the resulting algorithm to have a verifiable label complexity Λ such that $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$ is at most a $\mathrm{polylog}(1/(\varepsilon\delta))$ factor larger than the number of labels that would be requested by running $\mathcal{A}(\varepsilon, \delta)$ directly.

Similarly, given a budget-based active learning algorithm $\mathcal{A}$ having verifiable label complexity Λ, and given any values $\varepsilon, \delta \in (0, 1)$, we can produce a self-verifying algorithm by running $\mathcal{A}(2^i)$ for increasing values of $i \in \mathbb{N}$ until $\hat{\varepsilon}_{2^i, \delta/(i+1)^2} \leq \varepsilon$. With probability at least $1 - \delta$, the total number of labels requested by this self-verifying algorithm should be at most $\min\{2^{i+1} : 2^i \geq \Lambda(\varepsilon, \delta/(i+1)^2, \mathcal{P}_{XY})\}$ (or $\min\{2^{i+1} : 2^i \geq \Lambda(\nu + \varepsilon, \delta/(i+1)^2, \mathcal{P}_{XY})\}$ in the nonrealizable case); again, we expect this would typically differ only by logarithmic factors from $\Lambda(\varepsilon, \delta, \mathcal{P}_{XY})$ (or $\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY})$ in the nonrealizable case).

8.8 Classes of Infinite VC Dimension

So far, most of this article has focused on hypothesis classes $\mathbb{C}$ with finite VC dimension d (or finite pseudo-dimension in Chapter 6). However, there are also several results known for learning problems in which $d = \infty$, where the expressiveness of $\mathbb{C}$ is described by other notions of complexity.

One interesting such notion, commonly used in the passive learning literature, is the *uniform entropy*. Specifically, recalling the definition of $\mathcal{N}(\varepsilon, \mathcal{P})$, the ε-covering number of $\mathbb{C}$, from Chapter 4, we say $\mathbb{C}$ satisfies the uniform entropy condition, with values $\rho \in (0, 1)$ and $q \in [1, \infty)$, if

$$\forall \varepsilon > 0, \forall P, \mathrm{Log}(\mathcal{N}(\varepsilon, P)) \leq q\varepsilon^{-\rho}, \tag{8.9}$$

where P ranges over all finitely-discrete probability measures over $\mathcal{X}$.

A related notion of complexity, also commonly appearing in the passive learning literature, is the *bracketing entropy*. In this case, for classifiers g_1 and g_2, a *bracket* $[g_1, g_2]$ is the set of all classifiers g with $g_1(x) \leq g(x) \leq g_2(x)$ for all $x \in \mathcal{X}$. $[g_1, g_2]$ is called an ε-bracket under $L_1(\mathcal{P})$ if $\mathcal{P}(x : g_1(x) \neq g_2(x)) < \varepsilon/2$. Then, for any $\varepsilon > 0$, the value $\mathcal{N}_{[]}(\varepsilon, L_1(\mathcal{P}))$, called the *$\varepsilon$-bracketing number*, is defined as the smallest integer N such that there exist ε-brackets (under $L_1(\mathcal{P})$) $[g_{11}, g_{12}], \ldots, [g_{N1}, g_{N2}]$ with $\mathbb{C} \subseteq \bigcup_{i=1}^{N} [g_{i1}, g_{i2}]$: that is, the smallest number of ε-brackets sufficient to cover $\mathbb{C}$. If no such integer exists, define $\mathcal{N}_{[]}(\varepsilon, L_1(\mathcal{P})) = \infty$. We say $\mathbb{C}$ satisfies the bracketing entropy condition, with values $\rho \in (0, 1)$ and $q \in [1, \infty)$, if

$$\forall \varepsilon > 0, \mathrm{Log}(\mathcal{N}_{[]}(\varepsilon, L_1(\mathcal{P}))) \leq q\varepsilon^{-\rho}. \tag{8.10}$$

The following lemma results from combining various theorems from the passive learning literature [van der Vaart and Wellner, 1996, 2011, Koltchinskii, 2006, Giné and Koltchinskii, 2006].

Lemma 8.14. There is a universal constant $c \in (1, \infty)$ such that, if either (8.9) or (8.10) is satisfied with given values q and ρ, and Condition 2.3 is satisfied with given values a and α, and if we define

$$U'(m, \delta) = c \left(\frac{qa^{1-\rho}/(1-\rho)^2}{m} \right)^{\frac{1}{2-\alpha(1-\rho)}} + c \left(\frac{a\mathrm{Log}(1/\delta)}{m} \right)^{\frac{1}{2-\alpha}},$$

with probability at least $1-\delta$, $\forall h \in \mathbb{C}$, the following inequalities hold:

$$\mathrm{er}(h) - \mathrm{er}(f^\star) \leq \max\left\{2\left(\mathrm{er}_m(h) - \mathrm{er}_m(f^\star)\right), U'(m,\gamma)\right\},$$
$$\mathrm{er}_m(h) - \min_{g\in\mathbb{C}} \mathrm{er}_m(g) \leq \max\left\{2\left(\mathrm{er}(h) - \mathrm{er}(f^\star)\right), U'(m,\gamma)\right\}.$$

Based on this, there is a clear way to bound the label complexity of the $\mathrm{ERM}(\mathbb{C},\cdot)$ passive learning algorithm, simply by inverting the $U'(m,\delta)$ bound to obtain the smallest m for which $U'(m,\delta) \leq \varepsilon$. This bound is stated in the following classic theorem [see e.g., van der Vaart and Wellner, 1996, Mendelson, 2002].

Theorem 8.15. The passive learning algorithm $\mathrm{ERM}(\mathbb{C},\cdot)$ achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$, if either (8.9) or (8.10) is satisfied with given values q and ρ, and Condition 2.3 is satisfied with given values a and α, then for any $\varepsilon, \delta \in (0,1)$,

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim \frac{qa^{1-\rho}}{(1-\rho)^2}\left(\frac{1}{\varepsilon}\right)^{2-\alpha(1-\rho)} + a\left(\frac{1}{\varepsilon}\right)^{2-\alpha}\mathrm{Log}(1/\delta).$$

Since Lemma 8.14 has the same form as Lemma 3.1, it is reasonable to consider using the bound $U'(m,\delta)$ in place of $U(m,\delta)$ in RobustCAL. By essentially the same reasoning, this leads to an active learning method with label complexity sometimes smaller than that stated above for $\mathrm{ERM}(\mathbb{C},\cdot)$, as before multiplying it by roughly a factor of $\theta(a\varepsilon^\alpha)a\varepsilon^\alpha$ (aside from logarithmic factors). Formally, we have the following theorem; similar results for related methods have been obtained by Hanneke [2009a, 2011], Koltchinskii [2010], and Hanneke and Yang [2012].

Theorem 8.16. For any $\delta \in (0,1)$, if $\mathcal{P}_{XY}$ satisfies either (8.9) or (8.10) with given values q and ρ, and Condition 2.3 is satisfied with given values a and α, if we replace U with U' (from Lemma 8.14) in $\mathrm{RobustCAL}_\delta$, the resulting active learning algorithm achieves a label complexity Λ such that, for a constant $c \in (1,\infty)$ depending on α and ρ, $\forall \varepsilon \in (0,1)$,

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim \frac{cqa^{2-\rho}\theta(a\varepsilon^\alpha)}{(1-\rho)^2}\left(\frac{1}{\varepsilon}\right)^{2-\alpha(2-\rho)} + a^2\theta(a\varepsilon^\alpha)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\mathrm{Log}(1/\varepsilon).$$

As was the case with the original RobustCAL algorithm, it is possible to substitute a data-dependent estimator in place of $U'(m, \delta)$, so that the algorithm has no direct dependence on $\mathcal{P}_{XY}$, while maintaining the validity of Theorem 8.16. In fact, the exact *same* data-dependent estimator $\hat{U}(m, \delta)$ alluded to in Chapter 5 already suffices to achieve this result, so that no further modification to that algorithm is needed; the same is true of the related methods and analyses of Hanneke [2009a, 2011], Koltchinskii [2010], and Hanneke and Yang [2012]. Additionally, as was the case in Theorem 5.4, the logarithmic factors on the second term in this bound can be reduced in many cases; when $\alpha < 1$ (or $\theta(\varepsilon) = \Theta(\varepsilon^{-\kappa})$ for some $\kappa \in (0, 1]$), the final factor of $\mathrm{Log}(1/\varepsilon)$ can be replaced by a constant factor; furthermore, by a modification to the definition of δ_m, in this same case the factor of $\mathrm{Log}\left(\mathrm{Log}(a/\varepsilon)/\delta\right)$ can be replaced by $\mathrm{Log}(C/\delta)$, for a constant C [Hanneke and Yang, 2012].

8.8.1 Boundary Fragment Classes

Of course, Theorem 8.16 is only interesting if there are interesting learning problems with $\theta(\varepsilon) = o(1/\varepsilon)$ satisfying these entropy conditions. One such problem, which has been studied in detail in both the passive learning and active learning literatures [see van der Vaart and Wellner, 1996, Castro and Nowak, 2008, Wang, 2011], is the problem of learning smooth *boundary fragment* classes. Formally, in this learning scenario, there is a $k \in \mathbb{N}$ and $\gamma \in (k, \infty)$ such that $\mathcal{X} = [0, 1]^{k+1}$, and $\mathbb{C}$ is the set of classifiers f for which $\exists g : [0, 1]^k \to \mathbb{R}$ such that $\forall x_1, \ldots, x_{k+1} \in \mathbb{R}, f(x_1, \ldots, x_{k+1}) = \mathbb{1}^{\pm}_{[g(x_1,\ldots,x_k),\infty)}(x_{k+1})$, and g has partial derivatives up to order $\underline{\gamma} = \min\{n \in \mathbb{N} \cup \{0\} : n < \gamma\}$ all uniformly bounded by a constant, and with partial derivatives of total order $\underline{\gamma}$ that are Hölder continuous with exponent $\gamma - \underline{\gamma}$. In other words, for a fixed $x_1, \ldots, x_k \in \mathbb{R}$, the function $f(x_1, \ldots, x_k, \cdot)$ defines a *threshold* classifier, and the location of the threshold varies smoothly in $x_1, \ldots, x_k$. Functions of the type g described above are said to be *γ-order smooth on* $[0, 1]^k$.

It is known that, as long as $\mathcal{P}$ has a bounded density, this scenario satisfies (8.10) with $\rho = k/\gamma$ and a value of q depending on γ, k, the bound on derivatives, the coefficient from the Hölder condition, and the density bound [see van der Vaart and Wellner, 1996]. In particular, combined with Theorem 8.15, this means $\mathrm{ERM}(\mathbb{C}, \cdot)$ achieves a label complexity Λ with asymptotic dependence of $\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY})$ on ε of $O\left((1/\varepsilon)^{2-\alpha(1-k/\gamma)}\right)$. A result of this type was originally established by Tsybakov [2004] for a different passive learning algorithm; Tsybakov [2004] also proves a lower bound on the minimax label complexity of passive learning for this class, matching the dependence on ε in this bound; specifically, he shows that for any expected-error label complexity Λ achieved by a passive learning algorithm, and any values $a \geq 1$ and $\alpha \in [0, 1]$, $\sup_{\mathcal{P}_{XY}} \Lambda(\nu + \varepsilon, \mathcal{P}_{XY}) = \Omega\left((1/\varepsilon)^{2-\alpha(1-k/\gamma)}\right)$, where $\mathcal{P}_{XY}$ ranges over all distributions satisfying Condition 2.3 with the given values a and α.

Wang [2011] has obtained a bound on the disagreement coefficient for the problem of smooth boundary fragment classes. Specifically, he has proven that in this case, if $\mathcal{P}$ has a density that is always within a constant factor of some γ-order smooth function on $[0, 1]^{k+1}$, then $\forall h \in \mathbb{C}$,

$$\theta_h(\varepsilon) = O\left((1/\varepsilon)^{\frac{k}{\gamma+k}}\right). \tag{8.11}$$

Plugging this into Theorem 8.16, we see that RobustCAL achieves a label complexity Λ with asymptotic dependence of $\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY})$ on ε of

$$O\left((1/\varepsilon)^{2-\alpha\left(2-\frac{k}{\gamma}-\frac{k}{\gamma+k}\right)}\right).$$

When $\alpha > 0$ and γ is large, the represents an improvement over the result stated above for passive learning. Wang [2011] also proves a lower bound to complement (8.11), showing that if $\mathcal{P}$ has a bounded density that is also bounded away from 0, $\exists h \in \mathbb{C}$ such that

$$\theta_h(\varepsilon) = \Omega\left((1/\varepsilon)^{\frac{k}{\gamma+k}}\right).$$

The problem of active learning with boundary fragment classes was also studied by Castro and Nowak [2008], who prove a lower bound on the minimax label complexity. Specifically, they show that for any expected-error label complexity Λ achieved by an active learning algorithm, and any values $a \geq 1$ and $\alpha \in [0, 1]$, $\sup_{\mathcal{P}_{XY}} \Lambda(\nu + \varepsilon, \mathcal{P}_{XY}) =$

$\Omega\left((1/\varepsilon)^{2-\alpha(2-k/\gamma)}\right)$, where $\mathcal{P}_{XY}$ ranges over all distributions satisfying Condition 2.3 with the given values a and α, having marginal distribution $\mathcal{P}$ a uniform distribution over $[0,1]^{k+1}$. In fact, they show this lower bound holds even if $\mathcal{P}_{XY}$ is restricted to a special class of distributions, with the property that Condition 2.3 is even satisfied (with the given values a and α) for every one of the threshold-learning subproblems specified by taking arbitrary fixed values $x_1,\ldots,x_k \in \mathbb{R}$, and then replacing $\mathcal{P}_{XY}$ with the conditional distribution of $(X,Y) \sim \mathcal{P}_{XY}$ given $X \in \{(x_1,\ldots,x_k,x) : x \in [0,1]\}$.

Given this stronger condition on $\mathcal{P}_{XY}$, Castro and Nowak [2008] additionally propose an active learning algorithm that nearly achieves the above lower bound. Specifically, they pick a set of k-tuples $(x_1,\ldots,x_k) \in [0,1]^k$ on a grid of a carefully-chosen resolution (depending on the values k, γ, and α), and for each of these they apply an active learning method to the corresponding threshold problem in the one-dimensional subspace $\{(x_1,\ldots,x_k,x) : x \in [0,1]\}$; by the above noise assumption, and the fact that threshold classifiers have disagreement coefficient at most 2, we know each of these subproblems can be learned with an expected-error label complexity $\tilde{O}((1/\varepsilon)^{2-2\alpha} \vee \mathrm{Log}(1/\varepsilon))$ (though Castro and Nowak use a different algorithm to achieve this same effect). By interpolating between the threshold values learned by these easy subproblems, Castro and Nowak [2008] are then able to construct an estimator of the boundary fragment function $f^\star$. They show that if the grid size and number of queries per subproblem are set appropriately, the resulting algorithm achieves an expected-error label complexity Λ with $\Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) = O((1/\varepsilon)^{2-\alpha(2-k/\gamma)}\mathrm{Log}(1/\varepsilon))$. This has a better dependence on ε (by a factor of $(1/\varepsilon)^{\alpha k/(\gamma+k)}/\mathrm{Log}(1/\varepsilon))$) compared to the label complexity given above for RobustCAL. It is not presently known whether this smaller label complexity is achievable by some algorithm under the more general family of $\mathcal{P}_{XY}$ distributions satisfying Condition 2.3 (or even (2.1)), with the assumption of $\mathcal{P}$ being a uniform distribution.

8.8.2 Surrogate Losses and Classes of Infinite Pseudo-dimension

Above, we have seen that the RobustCAL algorithm can also be applied to classes of infinite VC dimension, achieving label complexities that depend on certain entropy conditions satisfied by the learning problem. These same ideas apply equally well to the variant of RobustCAL discussed in Chapter 6, which uses a relaxation of the 0-1 loss to arrive at an algorithm that is often computationally easier to use. Here we briefly describe the label complexity guarantees that result from applying RobustCAL$^{\ell}$ with classes of infinite VC dimension. Throughout this subsection, we continue the notational conventions introduced in Chapter 6.

Before stating the results, we first generalize the entropy conditions of (8.9) and (8.10). For any set $\mathcal{H}$ of measurable functions mapping $\mathcal{X} \to \bar{\mathcal{Y}}$, any distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $\varepsilon > 0$, let $\mathcal{N}(\varepsilon, \ell \circ \mathcal{F}, L_2(P))$ denote the value of the smallest $N \in \mathbb{N}$ such that there exist functions $g_1, \ldots, g_N$ mapping $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with the property that, $\forall f \in \mathcal{H}$, $\min_{1 \leq i \leq N} \mathbb{E}[(g_i(X,Y) - \ell(f(X)Y))^2] < \varepsilon^2$, where $(X,Y) \sim P$; if no such N exists, define $\mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(P)) = \infty$. We say $\mathcal{H}$ and ℓ satisfy the uniform entropy condition, with values $\rho \in (0,1)$ and $q \in [1, \infty)$, if

$$\forall \varepsilon > 0, \forall P, \mathrm{Log}(\mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(P))) \leq q\varepsilon^{-2\rho}, \tag{8.12}$$

where P ranges over all finitely-discrete probability measures over $\mathcal{X} \times \mathcal{Y}$. Similarly, for any measurable functions g_1, g_2 mapping $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a *bracket* $[g_1, g_2]$ is the set of all functions $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}$, $g_1(x,y) \leq g(x,y) \leq g_2(x,y)$; $[g_1, g_2]$ is called an ε-bracket under $L_2(\mathcal{P}_{XY})$ if $\mathbb{E}[(g_1(X,Y) - g_2(X,Y))^2] < \varepsilon^2$, where $(X,Y) \sim \mathcal{P}_{XY}$. Let $\mathcal{N}_{[]}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathcal{P}_{XY}))$ denote the value of the smallest $N \in \mathbb{N}$ such that there exist ε-brackets (under $L_2(\mathcal{P}_{XY})$) $[g_{11}, g_{12}], \ldots, [g_{N1}, g_{N2}]$ with the property that $\{(x,y) \mapsto \ell(f(x)y) : f \in \mathcal{H}\} \subseteq \bigcup_{i=1}^{N} [g_{i1}, g_{i2}]$; if no such N exists, define $\mathcal{N}_{[]}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathcal{P}_{XY})) = \infty$. We say $\mathcal{H}$, ℓ, and $\mathcal{P}_{XY}$ satisfy the bracketing entropy condition, with values $\rho \in (0,1)$ and $q \in [1, \infty)$, if

$$\forall \varepsilon > 0, \mathrm{Log}(\mathcal{N}_{[]}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathcal{P}_{XY}))) \leq q\varepsilon^{-2\rho}. \tag{8.13}$$

There is an analogue of Lemma 6.6 that holds under these entropy conditions, stated in the following lemma. This result follows from a combination of several theorems from the literature [van der Vaart and Wellner, 1996, Koltchinskii, 2006, Giné and Koltchinskii, 2006].

Lemma 8.17. There is a constant $c \in (1, \infty)$ such that, for any probability measure P over $\mathcal{X} \times \mathcal{Y}$ satisfying Condition 6.2 with given values b and β, for any set $\mathcal{H}$ of measurable functions $f : \mathcal{X} \to \bar{\mathcal{Y}}$ with $f^{\star}_{P,\ell} \in \mathcal{H}$, any $\delta \in (0,1)$, and any $m \in \mathbb{N}$, if either (8.12) or (8.13) is satisfied with given values ρ and q, letting

$$U'_{\ell}(m,\delta) = c\left(\left(\frac{qb^{1-\rho}}{(1-\rho)^2 m}\right)^{\frac{1}{2-\beta(1-\rho)}} + \left(\frac{b\mathrm{Log}(1/\delta)}{m}\right)^{\frac{1}{2-\beta}}\right) \wedge \bar{\ell},$$

if $\mathcal{L} = \{(X'_1, Y'_1), \ldots, (X'_m, Y'_m)\} \sim P^m$, then with probability at least $1-\delta$, $\forall f \in \mathcal{H}$, the following inequalities hold:

$$\mathrm{R}_{\ell}(f;P) - \mathrm{R}_{\ell}(f^{\star}_{P,\ell};P) \le \max\left\{2\left(\mathrm{R}_{\ell}(f;\mathcal{L}) - \mathrm{R}_{\ell}(f^{\star}_{P,\ell};\mathcal{L})\right), U'_{\ell}(m,\delta)\right\},$$

$$\mathrm{R}_{\ell}(f;\mathcal{L}) - \inf_{g\in\mathcal{H}} \mathrm{R}_{\ell}(g;\mathcal{L}) \le \max\left\{2\left(\mathrm{R}_{\ell}(f;P) - \mathrm{R}_{\ell}(f^{\star}_{P,\ell};P)\right), U'_{\ell}(m,\delta)\right\}.$$

Again, this immediately leads to the following classic result on the label complexity of empirical ℓ-risk minimization [see e.g., van der Vaart and Wellner, 1996, 2011, Mendelson, 2002].

Theorem 8.18. The passive learning algorithm $\mathrm{ERM}_{\ell}(\mathcal{F}, \cdot)$ achieves a label complexity Λ such that, for any distribution $\mathcal{P}_{XY}$ satisfying Condition 6.2 with given values b and β, satisfying Condition 2.3 with given values a and α, and having $f^{\star}_{\ell} \in \mathcal{F}$, if either (8.12) or (8.13) is satisfied (with $\mathcal{H} = \mathcal{F}$) with given values ρ and q, $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim \left(\frac{qb^{1-\rho}}{(1-\rho)^2}\right)\left(\frac{1}{\Psi_{\ell}(\varepsilon)^{2-\beta(1-\rho)}}\right) + \left(\frac{b}{\Psi_{\ell}(\varepsilon)^{2-\beta}}\right)\mathrm{Log}\left(\frac{1}{\delta}\right).$$

As above, since Lemma 8.17 has the same form as Lemma 6.6, if we replace $U_{\ell}(m/2, \delta_m)$ with $U'_{\ell}(m/2, \delta_m)$ in $\mathrm{RobustCAL}^{\ell}_{\delta}$, the same reasoning used in the original proof of Theorem 6.8 still applies, and leads to the following bound on the label complexity (due to Hanneke and Yang, 2012).

Theorem 8.19. Suppose ℓ is classification-calibrated. For any $\delta \in (0,1)$, if we replace U_ℓ with U'_ℓ (from Lemma 8.17) in $\text{RobustCAL}_\delta^\ell$, the resulting active learning algorithm achieves a label complexity Λ such that, for any $\mathcal{P}_{XY}$ satisfying Condition 2.3 with given values a and α, satisfying Condition 6.2 with given values b and β, and with $f_\ell^\star \in \mathcal{F}$, if either (8.12) or (8.13) is satisfied (with $\mathcal{H} = \mathcal{F}$) with given values ρ and q, for a constant $c \in (1,\infty)$ depending on α, β, and ρ, $\forall \varepsilon \in (0,1)$,

$$\begin{aligned}\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim &\left(\frac{cqb^{1-\rho}}{(1-\rho)^2}\right)\left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}}\right)\\ &+\left(\frac{b\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)^{2-\beta}}\right)\mathrm{Log}\left(\frac{\mathrm{Log}(1/\Psi_\ell(\varepsilon))}{\delta}\right)\mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right).\end{aligned}$$

Again, it is possible to replace U'_ℓ with a data-dependent estimator, so that the algorithm has no direct dependence on $\mathcal{P}_{XY}$, while still satisfying Theorem 8.19 [see Hanneke and Yang, 2012].

Additionally, as was true of classes with $d_\ell < \infty$, in the case of a classification-calibrated loss ℓ satisfying Condition 6.3, one can show slightly stronger results. Specifically, following essentially similar reasoning as lead to Theorem 6.9, one can show that a slight modification of $\text{RobustCAL}_\delta^\ell$ (analogous to that discussed in Section 6.5.3, but using U'_ℓ in Step $6'$ instead of U_ℓ) achieves a label complexity Λ such that, for b and β as in Lemma 6.4, for any $\mathcal{P}_{XY}$ with $f_\ell^\star \in \mathcal{F}$ and satisfying Condition 2.3 for given values a and α, if (8.12) is satisfied (with $\mathcal{H} = \mathcal{F}$), then

$$\begin{aligned}\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY}) \lesssim &\left(\frac{cqb^{1-\rho}}{(1-\rho)^2}\right)\left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta(1-\rho)}\mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)\\ &+b\left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta}\mathrm{Log}\left(\frac{\mathrm{Log}(1/\Psi_\ell(\varepsilon))}{\delta}\right)\mathrm{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right),\end{aligned}$$

for a constant $c \in (0,\infty)$ depending on α, β, and ρ. Likewise, for the case of (8.13), based on arguments about the validity of (8.13) under appropriate conditional distributions (with appropriate modifications of the value q it is satisfied for) [see Hanneke and Yang, 2012], one can also show that, with an appropriate modification of Step 6 (a bit more involved than the above case), $\text{RobustCAL}_\delta^\ell$ can be made to achieve a label complexity Λ such that, for b and β as in Lemma 6.4, for any $\mathcal{P}_{XY}$ with $f_\ell^\star \in \mathcal{F}$ and satisfying Condition 2.3 for given values a and α, if (8.13) is satisfied (with $\mathcal{H} = \mathcal{F}$), then

$$\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \left(\frac{cqb^{1-\rho}}{(1-\rho)^2}\right)\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)^\rho \left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+(1-\beta)(1-\rho)} + b\left(\frac{\theta(a\varepsilon^\alpha)a\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} \operatorname{Log}\left(\frac{\operatorname{Log}(1/\Psi_\ell(\varepsilon))}{\delta}\right)\operatorname{Log}\left(\frac{1}{\Psi_\ell(\varepsilon)}\right),$$

for a constant $c \in (0,\infty)$ depending on α, β, and ρ. See the work of Hanneke and Yang [2012] for the formal details of these results; in particular, the method studied by Hanneke and Yang [2012] uses a sufficiently general variant of Step 6 so that no modification is necessary to realize this result, and furthermore uses data-dependent estimators to avoid any direct dependence on $\mathcal{P}_{XY}$. As with all of the above results for CAL and RobustCAL, the logarithmic factors in these bounds can be reduced in many cases [see Hanneke and Yang, 2012].

8.8.3 Smooth Regression Functions

In all of the above results on learning with surrogate losses, we have made the assumption that $f_\ell^\star \in \mathcal{F}$. We have essentially been treating this as something that is simply needed in order to guarantee that the algorithms based on the surrogate loss ℓ are consistent. We then showed that, for certain types of surrogate losses, we could obtain label complexity bounds somewhat similar to those obtained for analogous algorithms that directly optimize the 0-1 loss.

However, another possibility not accounted for in this analysis is that the assumption of $f_\ell^\star \in \mathcal{F}$ may sometimes restrict the set of allowed distributions $\mathcal{P}_{XY}$ to such an extent that the optimal label complexity is actually *smaller* than would be the case if $\mathcal{P}_{XY}$ were merely restricted to have $\text{sign}(f_\ell^\star) \in \mathbb{C} = \{\text{sign}(f) : f \in \mathcal{F}\}$, or if the label complexities were characterized purely in terms of properties of $\mathbb{C}$. This is a possibility explored by Audibert and Tsybakov [2007]. Interestingly, they find that when ℓ is the quadratic loss, and $\mathcal{F}$ is a Hölder class of functions, the label complexities achievable by a certain passive learning method under the assumption that $f^\star_\ell \in \mathcal{F}$ are indeed smaller than the known label complexities achievable under related assumptions on the Bayes optimal classifier $\mathrm{sign}(\eta(\cdot) - 1/2)$. Minsker [2012] extends these findings to the active learning setting, and similarly finds that a certain active learning method (somewhat related to RobustCAL$^\ell$), based on using the quadratic loss ℓ as a surrogate for the 0-1 loss, achieves a smaller label complexity under the assumption that $f^\star_\ell \in \mathcal{F}$ than would be indicated by results such as Theorem 8.16 which rely only on properties of $\mathbb{C}$ and $\mathrm{sign}(f^\star_\ell)$. These findings raise interesting questions about the tightness of the analysis of methods based on optimizing the 0-1 loss, under restricted conditions on η, and about the best approach to the design of learning methods when noise assumptions are expressed as explicit constraints on the form of the η function. Here we include a brief summary of the formal results, along with a high-level description of the algorithm of Minsker [2012] below.

First, we specify the function class $\mathcal{F}$ as the set of functions f mapping $\mathcal{X} = [0,1]^k$ to $\mathbb{R}$, which are $\underline{\gamma}$ times continuously differentiable (for some $\gamma > 0$, and where $\underline{\gamma} = \min\{n \in \mathbb{N} \cup \{0\} : n < \gamma\}$), and $\forall x, x' \in \mathcal{X}$, satisfy $|g(x') - T_x(x')| \leq K\|x - x'\|_\infty^\gamma$, where T_x is the Taylor polynomial of degree $\underline{\gamma}$ of f at the point x, and $K > 0$ is a constant. This function class $\mathcal{F}$ is known as the $(\gamma, K, [0,1]^k)$-*Hölder class of functions*.

Audibert and Tsybakov [2007] propose a passive learning algorithm based on thresholding a certain estimator for nonparametric regression, and prove that it achieves an expected-error label complexity Λ such that $\sup_{\mathcal{P}_{XY}} \Lambda(\nu + \varepsilon, \mathcal{P}_{XY}) = O\left((1/\varepsilon)^{(2+k/\gamma)(1-\alpha)}\right)$, where $\mathcal{P}_{XY}$ ranges over all distributions such that $f^\star_\ell \in \mathcal{F}$, with ℓ as the quadratic loss, (2.1) is satisfied with given values a and α, and the marginal distribution $\mathcal{P}$ over $\mathcal{X}$ is within a (finite, nonzero) constant factor of the uniform distribution (they in fact study somewhat more general conditions as well; see Audibert and Tsybakov, 2007). Note that the assumption of $f^\star_\ell \in \mathcal{F}$ is equivalent to saying that the *regression function* $x \mapsto \mathbb{E}[Y|X = x] = 2\eta(x) - 1$ is in the $(\gamma, K, [0,1]^k)$-Hölder class of functions. They further prove a matching lower bound (up to constant factors) on the minimax label complexity for this problem for the case of $\alpha\gamma/(1-\alpha) \leq k$: that is, in this case, for any expected-error label complexity Λ achieved by a passive learning algorithm, $\sup_{\mathcal{P}_{XY}} \Lambda(\nu+\varepsilon, \mathcal{P}_{XY}) = \Omega\left((1/\varepsilon)^{(2+k/\gamma)(1-\alpha)}\right)$, where $\mathcal{P}_{XY}$ ranges over the same family of distributions as in the upper bound.

Minsker [2012] studies this same scenario, except in the active learning setting, and with a somewhat more constrained function class $\mathcal{F}$; specifically, he takes $\mathcal{F}$ to be the set of functions f in the $(\gamma, K, [0,1]^k)$-Hölder class of functions having the property that, for any region A in a recursive dyadic partition of $[0,1]^k$, $\sup_{x\in A}(f(x) - \mathbb{E}[f(X)|X \in A])^2 \leq c\mathrm{Var}(f(X)|X \in A)$, for some constant $c > 0$. He proposes an active learning algorithm, which makes use of the quadratic loss ℓ in a manner quite similar to RobustCAL$^{\ell}$, except with a few interesting twists. One minor change is that, rather than doubling the number of unlabeled samples processed between updates to V, the algorithm of Minsker [2012] updates V every time the number of *label requests* is doubled. A more significant change is that, instead of using the function class $\mathcal{F}$, the algorithm starts by using a very simple function class $\mathcal{F}_0$ consisting of constant functions. Then, before updating V, it first has the opportunity to increase the complexity of the function class to some larger $\mathcal{F}_i$ set, where $\mathcal{F}_i$ is the set of piecewise-constant functions, constant within each of the 2^{ki} regions of the i^{th} level in a recursive dyadic partition of $\mathcal{X}$. It determines an appropriate value of i by a model selection technique, akin to the method of structural risk minimization [Vapnik, 1998], using the samples Q just requested (i.e., those in the region of disagreement of V). Upon selecting this class $\mathcal{F}_i$, the update to the space V effectively takes the subset of functions f in this $\mathcal{F}_i$ class that have $\mathrm{sign}(f(x)) = \mathrm{sign}(h(x))$ for every $x \in \mathcal{X} \setminus \mathrm{DIS}(V)$ and h in the previous V, and have $|f(x) - (2\hat{\eta}_i(x) - 1)| \leq \Delta_k$ for every $x \in \mathrm{DIS}(V)$, where $k-1$ is the number of updates to V so far, $\hat{\eta}_i$ is an estimator of η with $2\hat{\eta}_i - 1 \in \mathcal{F}_i$, and Δ_k is a carefully-chosen value; the estimator $\hat{\eta}_i$ is essentially the least-squares estimator of η on the queried points, within the class $\mathcal{F}_i$, with only a minor modification to how the normalization is done.

The idea is that, because of the additional assumption, a bound on the L_2 distance of $\hat{\eta}$ to η within each region of the partition defining $\mathcal{F}_i$ (which is something one can produce) can be converted into a bound on the L_∞ distance within that region, and we can use this bound to inform the update of the set V (hence the $|f(x) - (2\hat{\eta}_i(x)| \leq \Delta_k$ constraint). Thus, as long as these Δ_k bounds remain valid bounds on this local L_∞ loss of the $2\hat{\eta}_i - 1$ function, we will only include a region A from the partition defining $\mathcal{F}_i$ in the set $\mathcal{X} \setminus \mathrm{DIS}(V)$ if $|2\hat{\eta}_i - 1| > \Delta_k$ within A, in which case we are confident that $\mathrm{sign}(2\eta(x) - 1) = \mathrm{sign}(2\hat{\eta}_i - 1)$ for every $x \in A$.

The main question for deriving rates of convergence is then how much of the space $\mathcal{X}$ will have $|2\eta(x) - 1|$ sufficiently large within the respective region for x, so that $|2\hat{\eta}_i(x) - 1| > \Delta_k$ within that region, which would then cause the region to be removed from $\mathrm{DIS}(V)$ after the update. This is the point at which (2.1) becomes essential, because (in combination with the smoothness assumption) it indicates that there are only a few small regions of $\mathcal{X}$ that have η very close to $1/2$. For the other regions of the space, which can be picked out even with a relatively coarse partition defining $\mathcal{F}_i$ (i.e., even for relatively small values of i), η will be far from $1/2$, so that $|2\hat{\eta}_i(x) - 1| > \Delta_k$ can quickly be achieved for the points x in those regions. This means $\mathrm{DIS}(V)$ will focus on those few locations having points x with $\eta(x)$ close to $1/2$. This has a double effect. First, in an effort to capture the behavior of $2\eta - 1$ within these regions, the value of i will be driven higher, which (combined with the above reasoning) eventually causes $\mathrm{DIS}(V)$ to become increasingly focused around precisely those points where $\eta(x)$ crosses $1/2$. Second, it makes the problem of identifying the optimal classification for the points $x \in \mathrm{DIS}(V)$ less crucial, since even the optimal classifier has a relatively high error rate on these points.

By formalizing and quantifying this reasoning, Minsker [2012] proves that this algorithm achieves a label complexity Λ such that, if $\gamma \leq 1$ and $\gamma\alpha/(1-\alpha) \leq k$, then for any $\varepsilon, \delta \in (0,1)$, $\sup_{\mathcal{P}_{XY}} \Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) = O\left((1/\varepsilon)^{(2+k/\gamma)(1-\alpha)-\alpha}(\mathrm{Log}(1/\varepsilon))^{\frac{2\alpha}{1-\alpha}}\right)$, where $\mathcal{P}_{XY}$ ranges over all distributions such that $f^\star_\ell \in \mathcal{F}$, where ℓ is the quadratic loss, (2.1) is satisfied with given values a and α, and the marginal distribution $\mathcal{P}$ over $\mathcal{X}$ is within a (finite, nonzero) constant factor of the uniform distribution (though, as with Audibert and Tsybakov, 2007, the work of Minsker, 2012, also explores more general conditions). Aside from logarithmic factors, this is essentially an improvement over the above passive learning result of Audibert and Tsybakov [2007] by roughly a factor of ε^{α}. However, it is not clear whether this strong of an improvement is achievable without the additional assumption that allows this analysis to relate the L_2 and L_∞ losses. Minsker [2012] also complements this result with a lower bound on the minimax label complexity; specifically, he shows that for any expected-error label complexity Λ achieved by an active learning method, $\sup_{\mathcal{P}_{XY}} \Lambda(\nu + \varepsilon, \mathcal{P}_{XY}) = \Omega\left((1/\varepsilon)^{(2+k/\gamma)(1-\alpha)-\alpha}\right)$, where $\mathcal{P}_{XY}$ ranges over the same family of distributions as in the above upper bound. Since the upper bound can easily be converted into a bound on the expected-error label complexity of this algorithm, with essentially the same dependence on ε, we can conclude that this method achieves near-minimax performance for this family of distributions $\mathcal{P}_{XY}$.

References

K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56, 1998.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

M.-F. Balcan and S. Hanneke. Robust interactive learning. In *Proceedings of the 25*$^{\text{th}}$ *Conference on Learning Theory*, 2012.

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

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

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

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010.

J. L. Balcázar, J. Castro, and D. Guijarro. A new abstract combinatorial dimension for exact learning via queries. *Journal of Computer and System Sciences*, 64(1):2–21, 2002.

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26*$^{\text{th}}$ *International Conference on Machine Learning*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, 2010.

A. Beygelzimer, D. Hsu, N. Karampatziakis, J. Langford, and T. Zhang. Efficient active learning. In *Proceedings of the 28*$^{\text{th}}$ *International Conference on Machine Learning*, 2011.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

V. I. Bogachev. *Gaussian Measures*. American Mathematical Society, Mathematical Surveys and Monographs, Book 62, 1998.

R. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008.

R.M. Castro and R.D. Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83:71–102, 2011.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.

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

S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412: 1767–1781, 2011.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Conference on Learning Theory*, 2005.

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

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the* 23$^{\mathrm{rd}}$ *Conference on Learning Theory*, 2010.

O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, To Appear, 2012.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4):1306–1326, 1987.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

B. B. Eisenberg. *On the Sample Complexity of PAC-Learning using Random and Chosen Examples*. PhD thesis, Massachusetts Institute of Technology, 1992.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13:255–279, 2012.

V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. On agnostic learning of parities, monomials and halfspaces. *SIAM Journal on Computing*, 39(2): 606–645, 2009.

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

E. Friedman. Active learning for smooth problems. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007a.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, 2007b.

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22$^{\text{nd}}$ *Conference on Learning Theory*, 2009a.

S. Hanneke. *Theoretical Foundations of Active Learning.* PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009b.

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, To appear, 2014.

S. Hanneke and L. Yang. Negative results for active learning with convex losses. In *Proceedings of the* 13$^{\text{th}}$ *International Conference on Artificial Intelligence and Statistics*, 2010.

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100: 78–150, 1992.

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory A*, 69:217–232, 1995.

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994.

T. Hegedüs. Generalized teaching dimension and the query complexity of learning. In *Proceedings of the 8th Conference on Computational Learning Theory*, 1995.

D. Helmbold, R. Sloan, and M. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5:165–196, 1990.

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

M. Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, 2006.

A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46^{th} Annual IEEE Symposium on Foundations of Computer Science*, 2005.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

S. R. Kulkarni. On metric entropy, Vapnik-Chervonenkis dimension, and learnability for a class of distributions. Technical Report CICS-P-160, Center for Intelligent Control Systems, 1989.

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

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

S. Mahalanabis. A note on active learning for smooth problems. arXiv*:1103.3095*, 2011.

S. Mahalanabis. *Subset and Sample Selection for Graphical Models: Gaussian Processes, Ising Models and Gaussian Mixture Models*. PhD thesis, Department of Computer Science, Edmund A. Hajim School of Engineering & Applied Sciences, University of Rochester, Rochester, New York, 2012.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:1808–1829, 1999.

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1):67–90, 2012.

J. R. Munkres. *Topology.* Prentice Hall, 2^{nd} edition, 2000.

R. D. Nowak. Generalized binary search. In *Proceedings of the 46*th *Allerton Conference on Communication, Control, and Computing*, 2008.

R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12), 2011.

D. Pollard. *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics and American Statistical Association, 1990.

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems 24*, 2011.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

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

S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996.

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

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

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011.

H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal of Control*, 6(1):131–147, 1968.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.