# Teaching Dimension and the Complexity of Active Learning

Steve Hanneke

Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213 USA
shanneke@cs.cmu.edu

**Abstract.** We study the label complexity of pool-based active learning in the PAC model with noise. Taking inspiration from extant literature on Exact learning with membership queries, we derive upper and lower bounds on the label complexity in terms of generalizations of *extended teaching dimension*. Among the contributions of this work is the first nontrivial general upper bound on label complexity in the presence of persistent classification noise.

## 1 Overview of Main Results

In supervised machine learning, it is becoming increasingly apparent that well-designed interactive learning algorithms can provide valuable improvements over passive algorithms in learning performance while reducing the amount of effort required of a human annotator. In particular, there is presently much interest in the pool-based active learning setting, in which a learner can request the label of any example in a large pool of unlabeled examples. In this case, one crucial quantity is the number of label requests required by a learning algorithm: the *label complexity*. This quantity is sometimes significantly smaller than the sample complexity of passive learning. A thorough theoretical understanding of these improvements seems essential to fully exploit the potential of active learning.

In particular, active learning is formalized in the PAC model as follows. The pool of $m$ unlabeled examples are sampled i.i.d. according to some distribution $\mathcal{D}$. A binary label is assigned to each example by a (possibly randomized) oracle, but is hidden from the learner unless it requests the label. The *error rate* of a classifier $h$ is defined as the probability of $h$ disagreeing with the oracle on a fresh example $X \sim \mathcal{D}$. A learning algorithm outputs a classifier $\hat{h}$ from a *concept space* $\mathbb{C}$, and we refer to the infimum error rate over classifiers in $\mathbb{C}$ as the *noise rate*, denoted $\nu$. For $\epsilon, \delta, \eta \in (0, 1)$, we define the *label complexity*, denoted $\#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, \eta)$, as the smallest number $q$ such that there is an algorithm that outputs a classifier $\hat{h} \in \mathbb{C}$, and for sufficiently large $m$, for any oracle with $\nu \leq \eta$, with probability at least $1 - \delta$ over the sample and internal randomness, the algorithm makes at most $q$ label requests and $\hat{h}$ has error rate at most $\nu + \epsilon$.[1]

[1] Alternatively, if we know $q$ ahead of time, we can have the algorithm halt if it ever tries to make more than $q$ queries. The analysis is nearly identical in either case.

The careful reader will note that this definition does not require the algorithm to be successful if $\nu > \eta$, distinguishing this from the fully agnostic setting [1]; we discuss possible methods to bridge this gap in later sections.

Kulkarni [2] has shown that if there is no noise, and one is allowed arbitrary binary valued queries, then $O\left(\log N(\epsilon)\right) \leq O\left(d \log \frac{1}{\epsilon}\right)$ queries suffice to PAC learn, where $N(\epsilon)$ denotes the size of a minimal $\epsilon$-cover of $\mathbb{C}$ with respect to $\mathcal{D}$, and $d$ is the VC dimension of $\mathbb{C}$. This bound often has exponentially better dependence on $\frac{1}{\epsilon}$, compared to the sample complexity of passive learning. However, many binary valued queries are unnatural and difficult to answer in practice. One of the driving motivations for research on the label complexity of active learning is identifying, in a general way, which concept spaces and distributions allow us to obtain this exponential improvement using only label requests for examples in the unlabeled sample. A further question is whether such improvements can be sustained in the presence of classification noise. In this paper, we investigate these questions from the perspective of a general analysis.

On the subject of learning through interaction, there is a rich literature concerning the complexity of Exact learning with membership queries [3, 4]. The interested reader should consult the limpid survey by Angluin [4]. The essential distinction between that setting and the setting we are presently concerned with is that, in Exact learning, the learning algorithm is required to *identify* the oracle's actual target function, rather than *approximating* it with high probability; on the other hand, in the Exact setting there is no classification noise and the algorithm can ask for the label of *any* example. In a sense, Exact learning with membership queries is a limiting case of PAC active learning. As such, we may hope to draw inspiration from the extant work on Exact learning when formulating an analysis for the PAC setting.

To quantify $\#MQ(\mathbb{C})$, the worst-case number of membership queries required for Exact learning with concept space $\mathbb{C}$, Hegedüs [3] defines a quantity called the *extended teaching dimension* of $\mathbb{C}$, based on the *teaching dimension* of Goldman & Kearns [5]. Letting $t_0$ denote this quantity, Hegedüs proves that

$$\max\{t_0, \log_2 |\mathbb{C}|\} \leq \#MQ(\mathbb{C}) \leq t_0 \log_2 |\mathbb{C}|,$$

where the upper bound is achieved by a version of the Halving algorithm.

Inspired by these results, we generalize the extended teaching dimension to the PAC setting, adding dependences on $\epsilon$, $\delta$, $\eta$, and $\mathcal{D}$. Specifically, we define two quantities, $t$ and $\tilde{t}$, both of which have $t_0$ as a limiting case. We show that

$$\Omega\left(\max\left\{\frac{\eta^2}{\epsilon^2}, \tilde{t}, \log N(2\epsilon)\right\}\right) \leq \#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, \eta) \leq \tilde{O}\left(\left(\frac{\eta^2}{\epsilon^2}+1\right) t \log N(\epsilon/2)\right)$$

where $\tilde{O}$ hides factors logarithmic in $\frac{1}{\epsilon}, \frac{1}{\delta}$, and $d$. The upper bound is achieved by an active learning algorithm inspired by the Halving algorithm, which uses $\tilde{O}\left(d \frac{\eta+\epsilon}{\epsilon^2}\right)$ unlabeled examples. With these tools in hand, we analyze the label complexity of axis-aligned rectangles with respect to product distributions, showing improvements over known passive learning results in dependence on $\eta$ when positive examples are not too rare.

The rest of the paper is organized as follows. In Section 2, we briefly survey the related literature on the label complexity of active learning. This is followed in Section 3 with the introduction of definitions and notation, and a brief discussion of known results for Exact learning in Section 4. In Section 5, we move into results for the PAC setting, beginning with the noise-free case for simplicity. Then, in Section 6, we describe the general setting, and prove an upper bound on the label complexity of active learning with noise; to the author's knowledge, this is the first general result of its kind, and along with lower bounds on label complexity presented in Section 7, represents the primary contribution of this work. We continue in Section 8, with an application of these bounds to describe the label complexity of axis-aligned rectangles with product distributions. We conclude with some enticing open problems in Section 9.

## 2 Context and Related Work

The recent literature studying general label complexity can be coarsely partitioned by the measure of progress used in the analysis. Specifically, there are at least three distinct ways to measure the progress of an active learning algorithm: *diameter* of the version space, *measure* of the region of disagreement, and *size* of the version space. By the *version space* at a time during the algorithm execution, we mean the set of concepts in $\mathbb{C}$ that have not yet been ruled out as a possible output. One approach to studying label complexity is to summarize in a single quantity how easy it is to make progress in terms of one of these progress metrics. This quantity, apart from itself being interesting, can then be used to derive upper and lower bounds on the label complexity.

To study the ease of reducing the diameter of the version space in active learning, Dasgupta [6] defines a quantity $\rho$ he calls the *splitting index.* $\rho$ is dependent on $\mathbb{C}$, $\mathcal{D}$, $\epsilon$, and another parameter $\tau$ he defines, as well as the oracle itself. Dasgupta finds that when the noise rate is zero, roughly $\tilde{O}(\frac{d}{\rho})$ label requests are sufficient, and $\Omega(\frac{1}{\rho})$ are necessary for learning (for respectively appropriate $\tau$ values). However, Dasgupta's analysis is restricted to the noise-free case, and there are no known extensions addressing the noisy case.

In studying ways to enable active learning in the presence of noise, Balcan et al. [1] propose the $A^2$ algorithm. This algorithm is able to learn in the presence of arbitrary classification noise. The strategy behind $A^2$ is to induce confidence intervals for the differences of error rates of concepts in the version space. If an estimated difference is statistically significant, the algorithm removes the worst of the two concepts. The key observation is that, since the algorithm only estimates error *differences*, there is no need to request the label of any example that all remaining concepts agree on. Thus, the number of label requests made by $A^2$ is largely controlled by how quickly the *region of disagreement* collapses as the algorithm progresses. However, apart from fall-back guarantees and a few special cases, there is presently no published general analysis of the number of label requests made by $A^2$, and no general index of how easy it is to reduce the region of disagreement.

The third progress metric is reduction in the *size* of the version space. If the concept space is infinite, an $\epsilon'$-cover of $\mathbb{C}$ can be substituted for $\mathbb{C}$, for some suitable $\epsilon'$.[2] This paper presents the first general study of the ease of reducing the size of the version space. The corresponding index summarizing the potential for progress in this metric remains informative in the presence of noise, given access to an upper bound on the noise rate.

In addition to the above studies, Kääriäinen [7] presents an interesting analysis of active learning with various types of noise. Specifically, he proves that under noise that is not persistent (in that requesting the same label twice may yield different responses) and where the Bayes optimal classifier is in $\mathbb{C}$, any algorithm that is successful for the zero noise setting can be transformed into a successful algorithm for the noisy setting with only a small increase in the number of label requests. However, these positive results do not carry into our present setting (*arbitrary persistent* classification noise). In fact, in addition to these positive results, Kääriäinen [7] presents negative results in the form of a general lower bound on the label complexity of active learning with arbitrary (persistent) classification noise. Specifically, he finds that for most nontrivial distributions $\mathcal{D}$, one can force any algorithm to make $\Omega\left(\frac{\nu^2}{\epsilon^2}\right)$ label requests.

## 3 Notation

We begin by introducing some notation. Let $\mathcal{X}$ be a set, called the *instance space*, and $\mathcal{F}$ be a corresponding $\sigma$-algebra. Let $\mathcal{D}_{XY}$ be a probability measure on $\mathcal{X}\times\{-1,1\}$. We use $\mathcal{D}$ to denote the marginal distribution of $\mathcal{D}_{XY}$ over $\mathcal{X}$. $\mathbb{C}_\mathcal{F}$ is the set of all $\mathcal{F}$-measurable $f : \mathcal{X} \to \{-1,1\}$. $\mathbb{C} \subseteq \mathbb{C}_\mathcal{F}$ is a concept space on $\mathcal{X}$, and we use $d$ to denote the VC dimension of $\mathbb{C}$; to focus on nontrivial learning, we assume $d > 0$. For any $h, h' \in \mathbb{C}_\mathcal{F}$, define $er_\mathcal{D}(h, h') = \mathcal{P}r_{X\sim\mathcal{D}}\{h(X) \neq h'(X)\}$. If $\mathcal{U} \in \mathcal{X}^m$, define $er_\mathcal{U}(h, h') = \frac{1}{m}\sum_{x\in\mathcal{U}} I[h(x) \neq h'(x)]$.[3] If $\mathcal{L} \in (\mathcal{X}\times\{-1,1\})^m$, define $er_\mathcal{L}(h) = \frac{1}{m}\sum_{(x,y)\in\mathcal{L}} I[h(x) \neq y]$. For any $h \in \mathbb{C}_\mathcal{F}$, define $er(h) = \mathcal{P}r_{(X,Y)\sim\mathcal{D}_{XY}}\{h(X) \neq Y\}$. Define the *noise rate* $\nu = \inf_{h\in\mathbb{C}} er(h)$. An $\alpha$-cover of $\mathbb{C}$ is any $V \subseteq \mathbb{C}$ s.t. $\forall h \in \mathbb{C}$, $\exists h' \in V$ with $er_\mathcal{D}(h, h') \leq \alpha$.

Generally, in this setting data is sampled i.i.d. according to $\mathcal{D}_{XY}$, but the labels are hidden from the learner unless it asks the oracle for them individually. In particular, requesting the same example's label twice gives the same label both times (though if the data sequence contains two identical examples, requesting

---

2 An alternative, but very similar progress metric is the size of an $\epsilon$-cover of the version space. The author suspects the analysis presented in this paper can be extended to describe that type of progress as well.

3 We overload the standard set-theoretic notation to also apply to sequences. In particular, $\sum_{x\in\mathcal{U}}$ indicates a sum over entries of the sequence $\mathcal{U}$ (not necessarily all distinct). Similarly, we use $|\mathcal{U}|$ to denote length of the sequence $\mathcal{U}$, $S \subseteq \mathcal{U}$ to denote a subsequence of $\mathcal{U}$, $S \cup \mathcal{U}$ to denote concatenation of two sequences, and for any particular $x \in \mathcal{U}$, $\mathcal{U} \setminus \{x\}$ indicates the subsequence of $\mathcal{U}$ with all entries except the single occurrence of $x$ that is implicitly referenced in the statement. It may help to think of each instance $x$ in a sample as having a unique identifier.

both labels might give two different values). However, for notational simplicity, we often abuse this notation by stating that $X \sim \mathcal{D}$ and later stating that the algorithm requests the label of $X$, denoted $Oracle(X)$; by this, we implicitly mean that $(X,Y) \sim \mathcal{D}_{XY}$, and the oracle reveals the value of $Y$ upon request. In particular, for $\mathcal{U} \sim \mathcal{D}^m$, $h \in \mathbb{C}_{\mathcal{F}}$, define $er_{\mathcal{U}}(h) = \frac{1}{m}\sum_{x\in\mathcal{U}} I[h(x) \neq Oracle(x)]$.

**Definition 1.** *For $V \subseteq \mathbb{C}$ with finite $|V|$, the* majority vote *concept $h_{maj} \in \mathbb{C}_{\mathcal{F}}$ is defined by $h_{maj}(x) = 1$ iff $|\{h \in V : h(x) = 1\}| \geq \frac{1}{2}|V|$.*

**Definition 2.** *For $\mathcal{U} \in \mathcal{X}^m$, $h \in \mathbb{C}_{\mathcal{F}}$, we overload notation to define the sequence of labels $h(\mathcal{U}) = \{h(x)\}_{x\in\mathcal{U}}$ assigned to entries of $\mathcal{U}$ by $h$. For $V \subseteq \mathbb{C}_{\mathcal{F}}$, $V[\mathcal{U}]$ denotes any subset of $V$ such that $\forall h \in V, |\{h' \in V[\mathcal{U}] : h'(\mathcal{U}) = h(\mathcal{U})\}| = 1$. $V[\mathcal{U}]$ represents the labelings of $\mathcal{U}$ realizable by $V$.*

## 4 Extended Teaching Dimension

**Definition 3.** *(Extended Teaching Dimension [3]) Let $V \subseteq \mathbb{C}$, $m \geq 0$, $\mathcal{U} \in \mathcal{X}^m$.*

$$\forall f \in \mathbb{C}_{\mathcal{F}},\ XTD(f,V,\mathcal{U}) = \inf\{t | \exists R \subseteq \mathcal{U} : |\{h \in V : h(R) = f(R)\}| \leq 1 \wedge |R| \leq t\}.$$

$$XTD(V,\mathcal{U}) = \sup_{f\in\mathbb{C}_{\mathcal{F}}} XTD(f,V,\mathcal{U}).$$

*For a given $f$, we call any $R \subseteq \mathcal{U}$ such that $|\{h \in V : h(R) = f(R)\}| \leq 1$ a* specifying set *for $f$ on $\mathcal{U}$ with respect to $V$.*[4]

The goal of Exact learning with membership queries is to ask for the labels $f(x)$ of individual examples $x \in \mathcal{X}$ until the only concept in $\mathbb{C}$ consistent with the observed labels is the target $f \in \mathbb{C}$. Hegedüs [3] presents the following algorithm.

```
Algorithm: MembHalving
Output: The target concept f ∈ ℂ
0. V ← ℂ
1. Repeat until |V| = 1
2.      Let h_maj be the majority vote of V
3.      Let R ⊆ 𝒳 be a minimal specifying set for h_maj on 𝒳 with respect to V
4.      Ask for the label f(x) of every x ∈ R
5.      Let V ← {h ∈ V | ∀x ∈ R, f(x) = h(x)}
6. Return the remaining element of V
```

**Theorem 1.** *(Exact Learning: Hegedüs [3]). Letting $\#MQ(\mathbb{C})$ denote the Exact learning query complexity of $\mathbb{C}$ with membership queries on any examples in $\mathcal{X}$, and $t_0 = XTD(\mathbb{C},\mathcal{X})$, then the following inequalities are valid if $|\mathbb{C}| > 2$.*

$$\max\{t_0, \log_2 |\mathbb{C}|\} \leq \#MQ(\mathbb{C}) \leq t_0 \log_2 |\mathbb{C}|.$$

*Furthermore, this upper bound is achieved by the MembHalving algorithm.*[5]

[4] We also overload all of these definitions in the obvious way for *sets* $\mathcal{U} \subseteq \mathcal{X}$.

[5] By a slight alteration to choose queries in a particular greedy order, Hegedüs is able to reduce this upper bound to $2\frac{t_0}{\log_2 t_0}\log_2|\mathbb{C}|$. However, it is the simpler form of the algorithm (presented here) that we draw inspiration from in the following sections.

The upper bound of Theorem 1 is clear when we view MembHalving as a version of the Halving algorithm [8]. That is, querying all examples in a specifying set for $h$ guarantees either $h$ makes a mistake or we identify $f$. Thus, querying a specifying set for $h_{maj}$ guarantees that we at least halve the version space.

The following definitions represent natural extensions of $XTD$ to the PAC setting. The relation of these quantities to the complexity of active learning is our primary focus.

**Definition 4.** *(XTD Growth Function) For $m \geq 0$, $V \subseteq \mathbb{C}$, $\delta \in [0,1]$,*

$$\begin{aligned} XTD(V,\mathcal{D},m,\delta) &= \inf\{t|\forall f \in \mathbb{C}_{\mathcal{F}},\ \mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^m}\{XTD(f,V[\mathcal{U}],\mathcal{U}) > t\} \leq \delta\}. \\ XTD(V,m) &= \sup_{\mathcal{U}\in\mathcal{X}^m} XTD(V[\mathcal{U}],\mathcal{U}). \end{aligned}$$

$XTD(\mathbb{C},\mathcal{D},m,\delta)$ plays an important role in distribution-dependent bounds on the label complexity, while $XTD(\mathbb{C},m)$ plays an analogous role in distribution-free bounds. Clearly $0 \leq XTD(\mathbb{C},\mathcal{D},m,\delta) \leq XTD(\mathbb{C},m) \leq m$.

As a simple example, consider the space of thresholds on the line. That is, suppose $\mathcal{X} = \mathbb{R}$ and $\mathbb{C} = \{h_\theta : \theta \in \mathbb{R}, h_\theta(x) = +1 \text{ iff } x \geq \theta\}$. In this case, $XTD(\mathbb{C},m) = 2$, since for any set $\mathcal{U}$ of $m$ points, and any $f \in \mathbb{C}_{\mathcal{F}}$, we can form a specifying set with the points $\min\{x \in \mathcal{U} : f(x) = +1\}$ and $\max\{x \in \mathcal{U} : f(x) = -1\}$, (if they exist).

## 5 The Complexity of Realizable Active Learning

Before discussing the general setting, we begin with realizable learning ($\eta = 0$), because the analysis is quite simple, and clearly highlights the relationship to the MembHalving algorithm. We handle noisy labels in the next section.

Based on Theorem 1, it should be clear that for $m \geq \Omega\left(\frac{1}{\epsilon}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$, $\#LQ(\mathbb{C},\mathcal{D},\epsilon,\delta,0) \leq XTD(\mathbb{C},m)d\log_2\frac{em}{d}$. Roughly speaking, this is achieved by drawing $m$ unlabeled examples $\mathcal{U}$ and executing MembHalving with concept space $\mathbb{C}[\mathcal{U}]$ and instance space $\mathcal{U}$. This gives a *data-dependent* bound of $XTD(\mathbb{C}[\mathcal{U}],\mathcal{U})\log_2|\mathbb{C}[\mathcal{U}]| \leq XTD(\mathbb{C},m)d\log_2\frac{em}{d}$. We can also obtain a related *distribution-dependent* result as follows. Consider the following algorithm.

```
Algorithm: ActiveHalving
Input: V ⊆ C_F, values ϵ, δ ∈ (0,1), U = {x_1, x_2, ..., x_m} ∈ X^m, constant n ∈ ℕ
Output: Concept ĥ ∈ V
0. Let i ← 0
1. Repeat
2.     i ← i + 1
3.     Let U_i = {x_{1+n(i-1)}, x_{2+n(i-1)}, ..., x_{ni}}
4.     Let h_maj be the majority vote of V
5.     Let R ⊆ U_i be a minimal specifying set for h_maj on U_i w.r.t. V[U_i]
6.     Ask for the label f(x) of every x ∈ R
7.     Let V ← {h ∈ V | f(R) = h(R)}
8.     If ∃h ∈ V s.t. h_maj(U_i) = h(U_i), Return argmin_{ĥ∈V} er_U(ĥ, h_maj)
```

**Theorem 2.** *Let* $m = \left\lceil \frac{256d}{\epsilon}\left(\ln \frac{92d}{\epsilon\delta}\right)^2 \right\rceil$*, and* $n = \left\lceil \frac{4}{\epsilon} \ln \frac{12d \log_2 \frac{4em}{\delta}}{\delta} \right\rceil$*. Let* $\hat{t} = XTD\left(\mathbb{C}, \mathcal{D}, n, \frac{\delta}{12d\log_2 \frac{4em}{\delta}}\right)$*. If* $N(\delta/(2m))$ *is the size of a minimal* $\frac{\delta}{2m}$*-cover of* $\mathbb{C}$*, then*

$$\#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, 0) \leq \hat{t}\log_2 N(\delta/(2m)) \leq O\left(\hat{t} d \log \frac{d}{\epsilon\delta}\right).$$

*Proof.* The bound is achieved by ActiveHalving$(V, \epsilon, \delta, \mathcal{U}, n)$, where $\mathcal{U} \sim \mathcal{D}^m$, and $V$ is a minimal $\frac{\delta}{2m}$-cover of $\mathbb{C}$. Let $f \in \mathbb{C}$ have $er(f) = 0$. Let $\hat{f} = \arg\min_{h\in V} er(h)$. With probability $\geq 1 - \delta/2$, $f(\mathcal{U}) = \hat{f}(\mathcal{U})$. Suppose this happens. In each iteration, if the condition in step 8 does not obtain, then either $\exists x \in R : h_{maj}(x) \neq f(x)$ or else $V[\mathcal{U}_i] = \{h\}$ for some $h \in V$ such that $\exists x \in \mathcal{U}_i : h_{maj}(x) \neq h(x) = f(x)$. Either way, we must have eliminated at least half of $V$ in step 7, so the condition in step 8 fails at most $\log_2 N(\delta/(2m)) < 2d\log_2 \frac{4em}{\delta} - 1$ times.

On the other hand, suppose the condition in step 8 obtains. This happens only when $h_{maj}(\mathcal{U}_i) = f(\mathcal{U}_i)$. $\mathcal{P}r_{\mathcal{U}_i}\left\{er_{\mathcal{U}_i}(h_{maj}, f) = 0 \wedge er_{\mathcal{U}}(h_{maj}, f) > \frac{\epsilon}{4}\right\} \leq \frac{\delta}{12d\log_2\frac{4em}{\delta}}$. By a union bound, the probability that an $h_{maj}$ with $er_{\mathcal{U}}(h_{maj}, f) > \frac{\epsilon}{4}$ satisfies the condition in step 8 on any iteration is at most $\frac{\delta}{6}$. If this does not happen, then the $\hat{h} \in V$ we return has $er_{\mathcal{U}}(\hat{h}, f) \leq er_{\mathcal{U}}(\hat{h}, h_{maj}) + er_{\mathcal{U}}(h_{maj}, f) \leq er_{\mathcal{U}}(f, h_{maj}) + er_{\mathcal{U}}(h_{maj}, f) \leq \frac{\epsilon}{2}$. By Chernoff and union bounds, $m$ is large enough so that with probability at least $1 - \frac{\delta}{6}$, $er_{\mathcal{U}}(\hat{h}, f) \leq \frac{\epsilon}{2} \Rightarrow er_{\mathcal{D}}(\hat{h}, f) \leq \epsilon$. So with probability $1 - \frac{5\delta}{6}$, we return an $\hat{h} \in \mathbb{C}$ with $er_{\mathcal{D}}(\hat{h}, f) \leq \epsilon$.

On the issue of number of queries, each iteration queries a minimal specifying set for $h_{maj}$ on a set of size $n$. The probability the size of this set is larger than $\hat{t}$ for a particular set $\mathcal{U}_i$ is at most $\frac{\delta}{12d\log_2\frac{4em}{\delta}}$. By a union bound, the probability it is larger than $\hat{t}$ on any iteration is at most $\frac{\delta}{6}$. Thus, the total probability of success (in learning and obtaining the query bound) is at least $1 - \delta$. □

Note that we can obtain a worst-case label bound for ActiveHalving by replacing $\hat{t}$ above with $XTD(\mathbb{C}, n)$. Theorem 2 highlights the relationship to known results in Exact learning with membership queries [3]. In particular, if $\mathbb{C}$ and $\mathcal{X}$ are finite, and $\mathcal{D}$ has support everywhere on $\mathcal{X}$, then as $\epsilon \to 0$ and $\delta \to 0$, the bound converges to $XTD(\mathbb{C}, \mathcal{X})\log_2 |\mathbb{C}|$, the upper bound in Theorem 1.

## 6 The Complexity of Active Learning with Noise

The following algorithm can be viewed as a noise-tolerant version of ActiveHalving. Significant care is needed to ensure we do not discard the best concept, and that the final classifier is near-optimal. The main trick is to use subsamples of size $< \frac{1}{16\eta}$. Since the probability of such a subsample containing a noisy example is small, the specifying sets for $h_{maj}$ will often be noise-free. Therefore, if $h \in V$ is contradicted in many such specifying sets, we can be confident $h$ is suboptimal. Likewise, if for a particular unqueried $x$, there are many such subsamples containing $x$ where $h_{maj}$ is *not* contradicted, and where there is a consistent $h$, then more often than not, $h(x) = h^*(x)$, where $h^* = \arg\min_{h'\in V} er(h')$.

Algorithm: $ReduceAndLabel(V, \mathcal{U}, \epsilon, \delta, \hat{\eta})$
Input: Finite $V \subseteq \mathbb{C}_{\mathcal{F}}$, $\mathcal{U} = \{x_1, x_2, \ldots, x_m\} \in \mathcal{X}^m$, values $\epsilon, \delta, \hat{\eta} \in (0, 1]$.
Output: Concept $h \in V$.
0. Let $u = \lfloor |\mathcal{U}| / (5 \ln |V|) \rfloor$
1. Let $V_0 \leftarrow V$, $i \leftarrow 0$
2. Do
3. $\quad i \leftarrow i + 1$
4. $\quad$ Let $\mathcal{U}_i = \{x_{1+u(i-1)}, x_{2+u(i-1)}, \ldots, x_{ui}\}$
5. $\quad V_i \leftarrow Reduce\left(V_{i-1}, \mathcal{U}_i, \frac{\delta}{48 \ln |V|}, \hat{\eta} + \frac{\epsilon}{2}\right)$
6. Until $|V_i| > \frac{3}{4}|V_{i-1}|$ or $|V_i| \leq 1$
7. Let $\bar{\mathcal{U}} = \{x_{ui+1}, x_{ui+2}, \ldots, x_{ui+\ell}\}$, where $\ell = \left\lceil 12 \frac{\hat{\eta}}{\epsilon^2} \ln \frac{12|V|}{\delta} \right\rceil$
8. $\mathcal{L} \leftarrow Label\left(V_{i-1}, \bar{\mathcal{U}}, \frac{\delta}{12}, \hat{\eta} + \frac{\epsilon}{2}\right)$
9. Return $h \in V_i$ having smallest $er_{\mathcal{L}}(h)$, (or any $h \in V$ if $V_i = \varnothing$)

Subroutine: $Reduce(V, \mathcal{U}, \delta, \hat{\eta})$
Input: Finite $V \subseteq \mathbb{C}_{\mathcal{F}}$, unlabeled sequence $\mathcal{U}$, values $\delta, \hat{\eta} \in (0, 1]$
Output: Concept space $V' \subseteq V$
0. Let $m = |\mathcal{U}|$, $n = \left\lfloor \frac{1}{16\hat{\eta}} \right\rfloor$, $r = \lceil 397 \ln \frac{2}{\delta} \rceil$, $\theta = \frac{27}{320}$
1. Let $h_{maj}$ be the majority vote of $V$
2. For $i \in \{1, 2, \ldots, r\}$
3. $\quad$ Sample a subsequence $S_i$ of size $n$ uniformly without replacement from $\mathcal{U}$
4. $\quad$ Let $R_i$ be a minimal specifying set for $h_{maj}$ in $S_i$ with respect to $V[S_i]$
5. $\quad$ Ask for the label of every example in $R_i$
6. $\quad$ Let $\bar{V}_i$ be the concepts $h \in V$ s.t. $h(R_i) \neq Oracle(R_i)$
7. Let $\bar{V}$ be the set of $h \in V$ that appear in $> \theta \cdot r$ of the sets $\bar{V}_i$
8. Return $V' = V \setminus \bar{V}$

Subroutine: $Label(V, \mathcal{U}, \delta, \hat{\eta})$
Input: Finite $V \subseteq \mathbb{C}_{\mathcal{F}}$, unlabeled sequence $\mathcal{U}$, values $\delta, \hat{\eta} \in (0, 1]$
Output: Labeled sequence $\mathcal{L}$
0. Let $\ell = |\mathcal{U}|$, $n = \left\lfloor \frac{1}{16\hat{\eta}} \right\rfloor$, $k = \lceil 167 \frac{\ell}{n} \ln \frac{3\ell}{\delta} \rceil$
1. Let $h_{maj}$ be the majority vote of $V$, and let $\mathcal{L} \leftarrow \{\}$
2. For $i \in \{1, 2, \ldots, k\}$
3. $\quad$ Sample a subsequence $S_i$ of size $n$ uniformly without replacement from $\mathcal{U}$
4. $\quad$ Let $R_i$ be a minimal specifying set for $h_{maj}$ in $S_i$ with respect to $V[S_i]$
5. $\quad$ For each $x \in R_i$, ask the oracle for its label $y_x$ and let $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y_x)\}$
6. Let $\hat{\mathcal{U}} \subseteq \mathcal{U}$ be the subsequence of examples we did not ask for the label of
7. For each $x \in \hat{\mathcal{U}}$
8. $\quad$ Let $\hat{I}_x = \{i : x \in S_i$ and $\exists h \in V$ s.t. $h(R_i) = h_{maj}(R_i) = Oracle(R_i)\}$
9. $\quad$ For each $i \in \hat{I}_x$, let $h_i \in V$ be s.t. $h_i(R_i) = Oracle(R_i)$
10. $\quad$ Let $y$ be the majority value of $\{h_i(x) : i \in \hat{I}_x\}$ (breaking ties arbitrarily)
11. $\quad$ Let $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y)\}$
12. Return $\mathcal{L}$

**Lemma 1.** *(Reduce) Suppose $h^* \in V$ is a concept such that $er_{\mathcal{U}}(h^*) \leq \hat{\eta} < \frac{1}{32}$. Let $V'$ be the set returned by Reduce$(V, \mathcal{U}, \epsilon, \delta, \hat{\eta})$. With probability at least $1-\delta$, $h^* \in V'$, and if $er_{\mathcal{U}}(h_{maj}, h^*) \geq 10\hat{\eta}$ then $|V'| \leq \frac{3}{4}|V|$.*

*Proof.* By a *noisy example*, in this context we mean any $x \in \mathcal{U}$ for which $h^*(x)$ disagrees with the oracle's label. Let $n = \left\lfloor \frac{1}{16\hat{\eta}} \right\rfloor$ and $r = \lceil 397 \ln \frac{2}{\delta} \rceil$, $\theta = \frac{27}{320}$. By a Chernoff bound, sampling $r$ subsequences of size $n$, each without replacement from $\mathcal{U}$, guarantees with probability $\geq 1 - \frac{\delta}{2}$ that at most $\theta r$ of the subsequences contain any noisy examples. In particular, this would imply $h^* \in V'$.

Now suppose $er_{\mathcal{U}}(h_{maj}, h^*) \geq 10\hat{\eta}$. For any particular subsampled sequence $S_i$, $\mathcal{P}r_{S_i \sim U_n(\mathcal{U})}\{h_{maj}(S_i) = h^*(S_i)\} \leq (1 - 10\hat{\eta})^n \leq 0.627$. So the probability there is some $x \in S_i$ with $h_{maj}(x) \neq h^*(x)$ is at least 0.373. By a Chernoff bound, with probability at least $1 - \frac{\delta}{2}$, at least $4\theta r$ of the $r$ subsamples contain some $x \in \mathcal{U}$ such that $h_{maj}(x) \neq h^*(x)$.

By a union bound, the total probability the above two events succeed is at least $1-\delta$. Suppose this happens. Any sequence $S_i$ containing no noisy examples but $\exists x \in S_i$ such that $h_{maj}(x) \neq h^*(x)$ necessarily has $|\bar{V}_i| \geq \frac{1}{2}|V|$. Since there are at least $3\theta r$ such subsamples $S_i$, we have $|\bar{V}| \geq \left(3\theta r \cdot \frac{1}{2}|V| - \theta r \cdot |V|\right) / (2\theta r) = \frac{1}{4}|V|$, so that $|V'| \leq \frac{3}{4}|V|$. □

**Lemma 2.** *(Label) Let $\mathcal{U} \in \mathcal{X}^\ell$, $\ell > n$. Suppose $h^* \in V$ has $er_{\mathcal{U}}(h^*) \leq \hat{\eta} < \frac{1}{32}$. Let $h_{maj}$ be the majority vote of $V$, and suppose $er_{\mathcal{U}}(h_{maj}, h^*) \leq 12\hat{\eta}$. Let $\mathcal{L}$ be the sequence returned by Label$(V, \mathcal{U}, \delta, \hat{\eta})$. With probability at least $1-\delta$, for every $(x, y) \in \mathcal{L}$, $y$ is either the oracle's label for $x$ or $y = h^*(x)$. In any case, $\forall x \in \mathcal{U}, |\{y : (x, y) \in \mathcal{L}\}| = 1$.*

*Proof.* As above, a *noisy example* is any $x \in \mathcal{U}$ such that $h^*(x)$ disagrees with the oracle. For any $x$ we ask for the label of, the entry $(x, y) \in \mathcal{L}$ has $y$ equal to the oracle's label, so the focus of the proof is on $\hat{\mathcal{U}}$. For each $x \in \hat{\mathcal{U}}$, let $I_x = \{i : x \in S_i\}$, $A = \{i : \exists x' \in R_i, h^*(x') \neq Oracle(x')\}$, and $B = \{i : \exists x' \in R_i, h_{maj}(x') \neq h^*(x')\}$. $\forall x \in \hat{\mathcal{U}}$, if $|I_x \cap A| < |(I_x \setminus B) \setminus A|$, we have that $|\{i \in I_x : h^*(R_i) = h_{maj}(R_i) = Oracle(R_i)\}| > \frac{1}{2}|\hat{I}_x| > 0$. In particular, this means the majority value of $\{h_i(x) : i \in \hat{I}_x\}$ is $h^*(x)$. The remainder of the proof bounds the probability this fails to happen.

For $x \in \hat{\mathcal{U}}$, for $i \in \{1, 2, \ldots, k\}$ let $\bar{S}_{i,x}$ of size $n$ be sampled uniformly without replacement from $\mathcal{U} \setminus \{x\}$, $\bar{A}_x = \{i : \exists x' \in \bar{S}_{i,x}, h^*(x') \neq Oracle(x')\}$, and $\bar{B}_x = \{i : \exists x' \in \bar{S}_{i,x}, h_{maj}(x') \neq h^*(x')\}$.

$$
\begin{aligned}
&\mathcal{P}r\left\{\exists x \in \hat{\mathcal{U}} : |I_x \cap A| \geq |(I_x \setminus B) \setminus A|\right\} \\
&\leq \sum_{x \in \mathcal{U}} \mathcal{P}r\left\{|I_x| < \tfrac{nk}{2\ell}\right\} + \mathcal{P}r\left\{|I_x \cap \bar{A}_x| \geq \tfrac{\sqrt{96}-1}{80}|I_x| \wedge |I_x| \geq \tfrac{nk}{2\ell}\right\} + \\
&\qquad \mathcal{P}r\left\{|(I_x \setminus \bar{B}_x) \setminus \bar{A}_x| \leq \tfrac{\sqrt{96}-1}{80}|I_x| \wedge |I_x| \geq \tfrac{nk}{2\ell}\right\} \leq \ell\left[e^{-\frac{kn}{8\ell}} + 2e^{-\frac{nk}{167\ell}}\right] \leq \delta.
\end{aligned}
$$

The second inequality is due to Chernoff and Hoeffding bounds. □

**Lemma 3.** *Suppose* $\nu = \inf_{h\in\mathbb{C}} er(h) \le \eta$ *and* $\eta + \frac{3}{4}\epsilon < \frac{1}{32}$. *Let* $V$ *be an* $\frac{\epsilon}{2}$*-cover of* $\mathbb{C}$. *Let* $\mathcal{U} \sim \mathcal{D}^m$, *with* $m = \left\lceil 224 \frac{\eta+\epsilon/2}{\epsilon^2} \ln \frac{48 \ln|V|}{\delta} \right\rceil \lceil 5 \ln |V| \rceil$. *Let* $n = \left\lfloor \frac{1}{16(\eta+3\epsilon/4)} \right\rfloor$, $\ell = \left\lceil 48 \frac{\eta+\epsilon/2}{\epsilon^2} \ln \frac{12|V|}{\delta} \right\rceil$, $s = \left\lceil 397 \ln \frac{96 \ln |V|}{\delta} \right\rceil (4 \ln |V|) + \left\lceil 167 \frac{\ell}{n} \ln \frac{36\ell}{\delta} \right\rceil$, *and* $t = XTD\left(V, \mathcal{D}, n, \frac{\delta}{2s}\right)$. *With probability* $\ge 1-\delta$, *ReduceAndLabel* $\left(V, \mathcal{U}, \frac{\epsilon}{2}, \delta, \eta + \frac{\epsilon}{2}\right)$ *makes at most* $ts$ *label queries and returns a concept* $h$ *with* $er(h) \le \nu + \epsilon$.

*Proof.* Let $h^* \in V$ have $er(h^*) \le \nu + \frac{\epsilon}{2}$. Suppose the value of $i$ is $\iota$ when we reach step 7. Clearly $\iota \le \log_{4/3} |V| \le 4 \ln |V|$. Let $h^i_{maj}$ denote the majority vote of $V_i$. We proceed by bounding the probability that any of six specific events fail to happen. The first event is

$$\left[\forall i \in \{1, 2, \ldots, \iota\}, er_{\mathcal{U}_i}(h^*) \le \eta + \tfrac{3}{4}\epsilon\right].$$

The probability this fails is $\le (4 \ln |V|) e^{-\left\lfloor \frac{m}{5 \ln |V|} \right\rfloor \frac{\epsilon^2}{\eta+\epsilon/2} \frac{1}{48}} \le \frac{\delta}{12}$ (by Chernoff and union bounds). The next event we consider is

$$\left[\forall i \in \{1, 2, \ldots, \iota\}, h^* \in V_i \text{ and (if } |V_\iota| > 1) \; er_{\mathcal{U}_\iota}\left(h^{\iota-1}_{maj}, h^*\right) < 10\left(\eta + \tfrac{3}{4}\epsilon\right)\right].$$

By Lemma 1 and a union bound, the previous event succeeds but this one fails with probability $\le \frac{\delta}{12}$. Next, note that the event

$$\left[\forall i \in \{1, 2, ..., \iota\}, er_{\mathcal{U}_i}\left(h^{i-1}_{maj}, h^*\right) < 10\left(\eta + \tfrac{3}{4}\epsilon\right) \Rightarrow er_{\mathcal{D}}\left(h^{i-1}_{maj}, h^*\right) \le \tfrac{21}{2}\left(\eta + \tfrac{3}{4}\epsilon\right)\right]$$

fails with probability $\le (4 \ln |V|) e^{-\left\lfloor \frac{m}{5 \ln |V|} \right\rfloor \left(\eta + \frac{3}{4}\epsilon\right) \frac{1}{84}} \le \frac{\delta}{12}$. The fourth event is

$$\left[er_{\bar{\mathcal{U}}}\left(h^{\iota-1}_{maj}, h^*\right) \le 12\left(\eta + \tfrac{3}{4}\epsilon\right)\right].$$

By a Chernoff bound, the probability this fails when the previous three events succeed is $\le e^{-\frac{\ell}{14}\left(\eta + \frac{3}{4}\epsilon\right)} \le \frac{\delta}{12}$. The fifth event is

$$[er_{\bar{\mathcal{U}}}(h^*) \le er(h^*) + \tfrac{\epsilon}{4} \text{ and } \forall h \in V_{\iota-1}, er(h) > er(h^*) + \tfrac{\epsilon}{2} \Rightarrow er_{\bar{\mathcal{U}}}(h) > er_{\bar{\mathcal{U}}}(h^*)].$$

By Chernoff and union bounds, the probability the previous events succeed but this fails is $\le |V| e^{-\frac{\ell}{48} \frac{\epsilon^2}{\eta+\epsilon/2}} \le \frac{\delta}{12}$. Finally, consider the event

$$[\forall (x, y) \in \mathcal{L}, y = h^*(x) \text{ or } y = Oracle(x)].$$

By Lemma 2, this fails when the other five succeed with probability $\le \frac{\delta}{12}$. Thus the probability all of these events succeed is $\ge 1 - \frac{\delta}{2}$. If they succeed, then any $h' \in V_\iota$ with $er(h') > \nu + \epsilon \ge er(h^*) + \frac{\epsilon}{2}$ has $er_{\mathcal{L}}(h') > er_{\mathcal{L}}(h^*) \ge \min_{h \in V_\iota} er_{\mathcal{L}}(h)$. Thus the $h$ we return has $er(h) \le \nu + \epsilon$.

In each call to *Reduce*, we ask for the labels of a minimal specifying set for $r = \left\lceil 397 \ln \frac{96 \ln |V|}{\delta} \right\rceil$ sequences of length $n$. For each, we make at most $t$ label requests with probability $\ge 1 - \frac{\delta}{2s}$, so the probability any call to *Reduce* makes more than $tr$ label requests is $\le \frac{4\delta r \ln |V|}{2s}$. Similarly, in *Label* we request the labels of a minimal specifying set for $\le k = \left\lceil 167 \frac{\ell}{n} \ln \frac{36\ell}{\delta} \right\rceil$ sequences of length $n$. So we make at most $tk$ queries in *Label* with probability $\ge 1 - \frac{\delta k}{2s}$. Thus, the total probability we make more than $t(k + 4r \ln |V|) = ts$ queries is $\le \frac{4\delta r \ln |V|}{2s} + \frac{\delta k}{2s} = \frac{\delta}{2}$. The total probability either the query or error bound is violated is at most $\delta$. □

**Theorem 3.** *Let* $n = \left\lfloor \frac{1}{16(\eta+3\epsilon/4)} \right\rfloor$, *and let* $N$ *be the size of a minimal* $\frac{\epsilon}{2}$*-cover of* $\mathbb{C}$. *Let* $\ell = \left\lceil 48\frac{\eta+\epsilon/2}{\epsilon^2} \ln \frac{12N}{\delta} \right\rceil$. *Let* $s = \left\lceil (397 \ln \frac{96 \ln N}{\delta}) \right\rceil (4 \ln N) + \left\lceil 167 \frac{\ell}{n} \ln \frac{36\ell}{\delta} \right\rceil$, *and* $t = XTD\left(\mathbb{C}, \mathcal{D}, n, \frac{\delta}{2s}\right)$.

$$\#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, \eta) \leq ts = O\left(t\left(\frac{\eta^2}{\epsilon^2}+1\right)\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\left(\log \frac{d}{\epsilon\delta}\right)\right).$$

*Proof.* It is known that $N < 2\left(\frac{4e}{\epsilon} \ln \frac{4e}{\epsilon}\right)^d$ [9]. If $\eta + \frac{3}{4}\epsilon \geq \frac{1}{32}$, the bound exceeds the passive sample complexity, so it clearly holds. Otherwise, the result follows from Lemma 3 and the fact that $XTD\left(V, \mathcal{D}, n, \frac{\delta}{2s}\right) \leq XTD\left(\mathbb{C}, \mathcal{D}, n, \frac{\delta}{2s}\right)$. □

Generally, if we do not know an upper bound $\eta$ on the noise rate $\nu$, then we can perform a guess-and-double procedure using a labeled validation set, which grows to size at most $\tilde{O}\left(\frac{\nu+\epsilon}{\epsilon^2}\right)$. See Section 9 for more discussion of this matter.

We can create a general algorithm, independent of $\mathcal{D}$, by using unlabeled examples to (with probability $\geq 1-\delta/2$) construct the $\frac{\epsilon}{2}$-cover. It is possible to do this while maintaining $|V| \leq N' = 2\left(\frac{16e}{\epsilon} \ln \frac{16e}{\epsilon}\right)^d$ using $O\left(\frac{1}{\epsilon^2}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$ unlabeled examples. Thus, replacing $t$ in Theorem 3 with $XTD(\mathbb{C}, n)$ and increasing $N$ to $N'$ gives an upper bound on the *distribution-free* label complexity.

## 7 Lower Bounds

In this section, we prove lower bounds on the label complexity.

**Definition 5.** *(Extended Partial Teaching Dimension) Let* $V \subseteq \mathbb{C}$, $m \geq 0$, $\delta \geq 0$. $\forall f \in \mathbb{C}_{\mathcal{F}}, \mathcal{U} \in \mathcal{X}^{\lceil m \rceil}$,

$$\begin{aligned}
XPTD(f, V, \mathcal{U}, \delta) &= \inf\{t | \exists R \subseteq \mathcal{U} : |\{h \in V : h(R) = f(R)\}| \leq \delta|V| + 1 \wedge |R| \leq t\}. \\
XPTD(V, \mathcal{D}, \delta) &= \inf\{t | \forall f \in \mathbb{C}_{\mathcal{F}}, \lim_{n\to\infty} \mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^n}\{XPTD(f, V, \mathcal{U}, \delta) > t\} = 0\}. \\
XPTD(V, m, \delta) &= \sup_{f\in\mathbb{C}_{\mathcal{F}}} \sup_{\mathcal{U}\in\mathcal{X}^{\lceil m \rceil}} XPTD(f, V[\mathcal{U}], \mathcal{U}, \delta).
\end{aligned}$$

**Theorem 4.** *Let* $\epsilon \in [0, 1/2)$, $\delta \in [0, 1)$. *For any* $2\epsilon$*-separated set* $V \subseteq \mathbb{C}$ *with respect to* $\mathcal{D}$,

$$\max\{\log\left[(1-\delta)|V|\right], XPTD(V, \mathcal{D}, \delta)\} \leq \#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, 0).$$

*If* $0 < \delta < 1/16$ *and* $0 < \epsilon/2 \leq \eta < 1/2$, *and there are* $h_1, h_2 \in \mathbb{C}$ *such that* $er_{\mathcal{D}}(h_1, h_2) = 2(\eta + \epsilon)$, *then*

$$\Omega\left(\left(\frac{\eta^2}{\epsilon^2}+1\right)\log\frac{1}{\delta}\right) \leq \#LQ(\mathbb{C}, \mathcal{D}, \epsilon, \delta, \eta).$$

*Also, the following distribution-free lower bound applies. If* $\forall x \in \mathcal{X}, \{x\} \in \mathcal{F}$,[6] *then letting* $\mathfrak{D}$ *denote the set of all probability distributions on* $\mathcal{X}$, *for* any $V \subseteq \mathbb{C}$,

$$XPTD(V, (1-\epsilon)/\epsilon, \delta) \leq \sup_{\mathcal{D}'\in\mathfrak{D}} \#LQ(\mathbb{C}, \mathcal{D}', \epsilon, \delta, 0).$$

[6] This condition is not necessary, but simplifies the proof.

*Proof.* The $\log [(1-\delta)|V|]$ lower bound is due to Kulkarni [2].

We prove the $XPTD(V, \mathcal{D}, \delta)$ lower bound by the probabilistic method as follows. If $\delta|V| + 1 \geq |V|$, the bound is trivially true, so assume $\delta|V| + 1 < |V|$ (and in particular, $|V| < \infty$). Let $m \geq 0$, $\tilde{t} = XPTD(V, \mathcal{D}, \delta)$. By definition of $\tilde{t}$, $\exists f' \in \mathbb{C}_{\mathcal{F}}$ such that $\lim_{n\to\infty} \mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^n}\{XPTD(f', V, \mathcal{U}, \delta) \geq \tilde{t}\} > 0$. By the Dominated Convergence Theorem and Kolmogorov's Zero-One Law, this implies $\lim_{n\to\infty} \mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^n}\{XPTD(f', V, \mathcal{U}, \delta) \geq \tilde{t}\} = 1$. Since this probability is nonincreasing in $n$, this means $\mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^m}\{XPTD(f', V, \mathcal{U}, \delta) \geq \tilde{t}\} = 1$. Suppose $\mathcal{A}$ is a learning algorithm. For $\mathcal{U} \in \mathcal{X}^m$, $f \in \mathbb{C}_{\mathcal{F}}$, define random quantities $R_{\mathcal{U},f} \subseteq \mathcal{U}$ and $h_{\mathcal{U},f} \in \mathbb{C}$, denoting the examples queried and classifier returned by $\mathcal{A}$, respectively, when the oracle answers consistent with $f$ and the input unlabeled sequence is $\mathcal{U} \sim \mathcal{D}^m$. If we sample $f$ uniformly at random from $V$,

$$
\begin{aligned}
&\mathbb{E}_f\left[\mathcal{P}r_{\mathcal{U},R_{\mathcal{U},f},h_{\mathcal{U},f}}\left\{er_{\mathcal{D}}(f, h_{\mathcal{U},f}) > \epsilon \vee |R_{\mathcal{U},f}| \geq \tilde{t}\right\}\right] \\
&\geq \mathcal{P}r_{f,\mathcal{U},R_{\mathcal{U},f},h_{\mathcal{U},f}}\left\{f(R_{\mathcal{U},f}) = f'(R_{\mathcal{U},f}) \wedge er_{\mathcal{D}}(f, h_{\mathcal{U},f}) > \epsilon \vee |R_{\mathcal{U},f}| \geq \tilde{t}\right\} \\
&\geq \mathbb{E}_{\mathcal{U}}\left[\inf_{h\in\mathbb{C}, R\subseteq\mathcal{U}:|R|<\tilde{t}} \mathcal{P}r_f\left\{f(R) = f'(R) \wedge er_{\mathcal{D}}(h, f) > \epsilon\right\}\right] > \delta.
\end{aligned}
$$

Therefore, there must be some fixed target $f \in \mathbb{C}$ such that the probability that $er_{\mathcal{D}}(f, h_{\mathcal{U},f}) > \epsilon$ or $|R_{\mathcal{U},f}| \geq XPTD(V, \mathcal{D}, \delta)$ is $> \delta$, proving the lower bound.

Kääriäinen [7] proves a distribution-free version of the $\Omega\left(\left(\frac{\eta^2}{\epsilon^2} + 1\right)\log\frac{1}{\delta}\right)$ bound, and also mentions its extendibility to the distribution-dependent setting. Since the distribution-dependent claim and proof thereof are only implicit in that reference, for completeness we present a brief proof here. Let $\Delta = \{x : h_1(x) \neq h_2(x)\}$. Suppose $h^*$ is chosen from $\{h_1, h_2\}$ by an adversary. Given $\mathcal{D}$, we construct a distribution $\mathcal{D}_{XY}$ with the following property[7]. $\forall A \in \mathcal{F}, \mathcal{P}r_{(X,Y)\sim\mathcal{D}_{XY}}\{Y = h^*(X)|X \in A \cap \Delta\} = \frac{1}{2} + \frac{\epsilon}{2(\eta+\epsilon)}$, and $\mathcal{P}r_{(X,Y)\sim\mathcal{D}_{XY}}\{Y = h_1(X)|X \in A\backslash\Delta\} = 1$. Any concept $h \in \mathbb{C}$ with $er(h) \leq \eta+\epsilon$ has $\mathcal{P}r\{h(X) = h^*(X)|h_1(X) \neq h_2(X)\} > \frac{1}{2}$. Since this probability can be estimated to arbitrary precision with arbitrarily high probability using unlabeled examples, we have a reduction to active learning from the task of determining with probability $\geq 1-\delta$ whether $h_1$ or $h_2$ is $h^*$. Examining the latter task, since every subset of $\Delta$ in $\mathcal{F}$ yields the same conditional distribution, any optimal strategy is based on samples from this distribution. It is known (e.g., [10, 11]) that this requires expected number of samples at least

$$
\frac{(1-8\delta)\log\frac{1}{8\delta}}{8D_{KL}\left(\frac{1}{2}+\frac{\epsilon}{2(\eta+\epsilon)}||\frac{1}{2}-\frac{\epsilon}{2(\eta+\epsilon)}\right)} > \frac{1}{40}\frac{(\eta+\epsilon)^2}{\epsilon^2}\ln\frac{1}{8\delta},
$$

where $D_{KL}(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

We prove the $XPTD(V, (1-\epsilon)/\epsilon, \delta)$ bound as follows. Let $n = \frac{1-\epsilon}{\epsilon}$. For $S \in \mathcal{X}^{\lceil n\rceil}$, let $\mathcal{D}_S$ be the uniform distribution on the entries of $S$. Let $f'' \in \mathbb{C}_{\mathcal{F}}$ be such that $XPTD(f'', V[S], S, \delta) = \max_{f\in V[S]} XPTD(f, V[S], S, \delta)$, and define

[7] Although this proof relies on stochasticity of the oracle, with additional assumptions on $\mathcal{D}$ and $\Delta$ similar to Kääriäinen's [7], a similar result holds for deterministic oracles.

$t'' = XPTD(f'', V[S], S, \delta)$. Let $m \geq 0$. Let $R_{\mathcal{U},f}$ and $h_{\mathcal{U},f}$ be defined as above, for $\mathcal{U} \sim \mathcal{D}_S^m$. As above, we use the probabilistic method, this time by sampling the target function $f$ uniformly from $V[S]$.

$$\begin{aligned}
&\mathbb{E}_f \left[ \mathcal{P}r_{\mathcal{U}, R_{\mathcal{U},f}, h_{\mathcal{U},f}} \left\{ er_{\mathcal{D}_S}(h_{\mathcal{U},f}, f) > \epsilon \vee |R_{\mathcal{U},f}| \geq t'' \right\} \right] \\
&\geq \mathbb{E}_{\mathcal{U}} \left[ \mathcal{P}r_{f, R_{\mathcal{U},f}, h_{\mathcal{U},f}} \left\{ f(R_{\mathcal{U},f}) = f''(R_{\mathcal{U},f}) \wedge h_{\mathcal{U},f}(S) \neq f(S) \vee |R_{\mathcal{U},f}| \geq t'' \right\} \right] \\
&\geq \min_{h \in \mathbb{C}, R \subseteq S : |R| < t''} \mathcal{P}r_f \left\{ f(R) = f''(R) \wedge h(S) \neq f(S) \right\} > \delta.
\end{aligned}$$

Taking the supremum over $S \in \mathcal{X}^{\lceil n \rceil}$ completes the proof. □

## 8 Example: Axis-Aligned Rectangles

As an application, we analyze axis-aligned rectangles, when $\mathcal{D}$ is a product density. An axis-aligned rectangle in $\mathbb{R}^n$ is defined by a sequence $\{(a_i, b_i)\}_{i=1}^n$, such that $a_i \leq b_i$, and the examples labeled $+1$ are $\{x \in \mathcal{X} : \forall i, a_i \leq x \leq b_i\}$. Throughout this section, we assume $\mathcal{F}$ is the standard Borel $\sigma$-algebra on $\mathbb{R}^n$.

**Lemma 4.** *(Balanced Axis-Aligned Rectangles) If $\mathcal{D}$ is a product distribution on $\mathbb{R}^n$ with continuous CDF, and $\mathbb{C}$ is the set of axis-aligned rectangles such that $\forall h \in \mathbb{C}, \mathcal{P}r_{X \sim \mathcal{D}}\{h(X) = +1\} \geq \lambda$, then*

$$XTD(\mathbb{C}, \mathcal{D}, m, \delta) \leq O\left( \frac{n^2}{\lambda} \log \frac{nm}{\delta} \right).$$

*Proof.* If $G_i$ is the CDF of $X_i$ for $X \sim \mathcal{D}$, then $G_i(X_i)$ is uniform in $(0,1)$, and for any $h \in \mathbb{C}$, the function $h'(x) = h(\{\min\{y : x_i = G_i(y)\}\}_{i=1}^n)$ (for $x \in (0,1)^n$) is an axis-aligned rectangle. This mapping of the problem into $(0,1)^n$ is equivalent to the original, so for the rest of this proof, we assume $\mathcal{D}$ is uniform on $(0,1)^n$.

If $m$ is smaller than the bound, the result clearly holds, so assume $m \geq 2n + \frac{4n}{\lambda}\left(\ln \frac{8n}{\delta} + 2n \ln \frac{2nm^2}{\delta}\right)$. Our first step is to discretize the concept space. Let $S$ be the set of concepts $h$ such that the region $\{x : h(x) = +1\}$ is specified by the interior of some rectangle $\{(a_i, b_i)\}_{i=1}^n$ with $a_i, b_i \in \left\{0, \frac{\delta}{2nm^2}, 2\frac{\delta}{2nm^2}, \ldots, \left\lceil \frac{2nm^2}{\delta} \right\rceil \frac{\delta}{2nm^2} \right\}$, $a_i < b_i$. By a union bound, with probability $\geq 1 - \delta/2$ over the draw of $\mathcal{U} \sim \mathcal{D}^m$, $\forall x, y \in \mathcal{U}$, $\forall i \in \{1, 2, \ldots, n\}$, $|x_i - y_i| > \frac{\delta}{2nm^2}$. In particular, this would imply there are valid choices of $S[\mathcal{U}]$ and $\mathbb{C}[\mathcal{U}]$ so that $\mathbb{C}[\mathcal{U}] \subseteq S[\mathcal{U}]$. As such, $XTD(\mathbb{C}, \mathcal{D}, m, \delta) \leq XTD(S \cap \mathbb{C}, \mathcal{D}, m, \delta/2)$.

Let $f \in \mathbb{C}_{\mathcal{F}}$. If $\mathcal{P}r_{X \sim \mathcal{D}}\{f(X) = +1\} < \frac{3}{4}\lambda$, then with probability $\geq 1 - \delta/2$, for each $h \in S \cap \mathbb{C}$, there is at least one $x \in \mathcal{U}$ s.t. $h(x) = +1 \neq f(x)$. Thus $\mathcal{P}r_{\mathcal{U} \sim \mathcal{D}^m}\left\{XTD(f, (\mathbb{C} \cap S)[\mathcal{U}], \mathcal{U}) > \frac{4}{\lambda} \ln \frac{2|S|}{\delta}\right\} \leq \delta/2$.

For any set of examples $R$, let $CLOS(R)$ be the smallest axis-aligned rectangle $h \in S$ that labels all of $R$ as $+1$. This is known as the *closure* of $R$. Additionally, let $A \subseteq R$ be a smallest set such that $CLOS(A) = CLOS(R)$. This is known as a minimal spanning set of $R$. Clearly $|A| \leq 2n$, since the extreme points in each direction form a spanning set.

Let $h \in S$ be such that $\mathcal{P}r_{X\sim\mathcal{D}}\{h(X) = +1\} \geq \frac{\lambda}{2}$. Let $\{(a_i, b_i)\}_{i=1}^n$ define the rectangle. Let $x^{(ai)}$ be the example in $\mathcal{U}$ with largest $x_i^{(ai)}$ component such that $x_i^{(ai)} < a_i$ and $\forall j \neq i, a_j \leq x_j^{(ai)} \leq b_j$, or if no such example exists, $x^{(ai)}$ is defined as the $x \in \mathcal{U}$ with smallest $x_i$. Let $x^{(bi)}$ be defined similarly, except having the smallest $x_i^{(bi)}$ component with $x_i^{(bi)} > b_i$, and again $\forall j \neq i, a_j \leq x_j^{(bi)} \leq b_j$. If no such example exists, then $x^{(bi)}$ is defined as the $x \in \mathcal{U}$ with largest $x_i$. Let $A_{h,\mathcal{U}} \subseteq \mathcal{U}$ be the subsequence of all examples $x \in \mathcal{U}$ such that $\exists i \in \{1, 2, \ldots, n\}$ with $x_i^{(ai)} \leq x_i < a_i$ or $b_i < x_i \leq x_i^{(bi)}$. The surface volume of each face of the rectangle is at least $\lambda/2$. By a union bound over the $2n$ faces of the rectangle, with probability at least $1 - \delta/(4|S|)$, $|A_{h,\mathcal{U}}| \leq \frac{4n}{\lambda} \ln \frac{8n|S|}{\delta}$. With probability $\geq 1 - \delta/4$, this is satisfied for every $h \in S$ with $\mathcal{P}r_{X\sim\mathcal{D}}\{h(X) = +1\} \geq \frac{\lambda}{2}$.

Now suppose $f \in \mathbb{C}_{\mathcal{F}}$ satisfies $\mathcal{P}r_{X\sim\mathcal{D}}\{f(X) = +1\} \geq \frac{3\lambda}{4}$. Let $\mathcal{U}_+ = \{x \in \mathcal{U} : f(x) = +1\}$, $h_{clos} = CLOS(\mathcal{U}_+)$. If any $x \in \mathcal{U} \setminus \mathcal{U}_+$ has $h_{clos}(x) = +1$, we can form a specifying set for $f$ on $\mathcal{U}$ with respect to $S[\mathcal{U}]$ using a minimal spanning set for $\mathcal{U}_+$ along with this $x$. If there is no such $x$, then $h_{clos}(\mathcal{U}) = f(\mathcal{U})$, and we use a minimal specifying set for $h_{clos}$. With probability $\geq 1-\delta/4$, for every $h \in S$ such that $\mathcal{P}r_{X\sim\mathcal{D}}\{h(X) = +1\} < \frac{\lambda}{2}$, there is some $x \in \mathcal{U}_+$ such that $h(x) = -1$. If this happens, since $h_{clos} \in S$, this implies $\mathcal{P}r_{X\sim\mathcal{D}}\{h_{clos}(X) = +1\} \geq \frac{\lambda}{2}$. In this case, for a specifying set, we use $A_{h_{clos},\mathcal{U}}$ along with a minimal spanning set for $\{x \in \mathcal{U} : f(x) = +1\}$. So $\mathcal{P}r_{\mathcal{U}\sim\mathcal{D}^m}\left\{XTD(f, (\mathbb{C} \cap S)[\mathcal{U}], \mathcal{U}) > 2n + \frac{4n}{\lambda} \ln \frac{8n|S|}{\delta}\right\}$ $\leq \delta/2$. Noting that $|S| \leq \left(\frac{2nm^2}{\delta}\right)^{2n}$ completes the proof. □

Note that we can obtain an estimate $\hat{p}$ of $p = \mathcal{P}r_{(X,Y)\sim\mathcal{D}_{XY}}\{Y = +1\}$ that, with probability $\geq 1 - \delta/2$, satisfies $p/2 \leq \hat{p} \leq 2p$, using at most $O\left(\frac{1}{p} \log \frac{1}{p\delta}\right)$ labeled examples (by guess-and-halve). Since clearly $\mathcal{P}r_{X\sim\mathcal{D}}\{h^*(X) = +1\} \geq p - \eta$, we can take $\lambda = (\hat{p}/2) - \eta$, giving the following oracle-dependent bound.

**Theorem 5.** *If $\mathcal{D}$ is as in Lemma 4 and $\mathbb{C}$ is the set of* all *axis-aligned rectangles, then if $p = \mathcal{P}r_{(X,Y)\sim\mathcal{D}_{XY}}\{Y = +1\} > 4\eta$, we can, with probability $\geq 1 - \delta$, find an $h \in \mathbb{C}$ with $er(h) \leq \nu + \epsilon$ without the number of label requests exceeding*

$$\tilde{O}\left(\frac{n^3}{(p/4) - \eta}\left(\frac{\eta^2}{\epsilon^2} + 1\right)\right).$$

This result is somewhat encouraging, since if $\eta < \epsilon$ and $p$ is not too small, the label bound represents an exponential improvement in $\frac{1}{\epsilon}$ compared to known results for passive learning, while maintaining polylog dependence on $\frac{1}{\delta}$ and polynomial dependence on $n$, though the degree increases from 1 to 3. We might wonder whether the property of being balanced is sufficient for these improvements. However, as the following theorem shows, balancedness alone is insufficient for guaranteeing polylog dependence on $\frac{1}{\epsilon}$. The proof is omitted for brevity.

**Theorem 6.** *If $n \geq 2$, there is a distribution $\mathcal{D}'$ on $\mathbb{R}^n$ such that, if $\mathbb{C}$ is the set of axis-aligned rectangles $h$ with $\mathcal{P}r_{X\sim\mathcal{D}'}\{h(X) = +1\} \geq \lambda$, then there is a $V \subset \mathbb{C}$ $2\epsilon$-separated with respect to $\mathcal{D}'$ such that $\Omega\left(\frac{(1-\delta)(1-\lambda)}{\epsilon}\right) \leq XPTD(V, \mathcal{D}', \delta)$.*

## 9 Open Problems

There are a number of possibilities for tightening these bounds. The upper bound of Theorem 3 contains a $O\left(\log \frac{d}{\epsilon\delta}\right)$ factor, which does not appear in any known lower bounds. In the worst case, when $XTD(\mathbb{C}, \mathcal{D}, n, \delta) = O(n)$, this factor clearly does not belong, since the bound exceeds the passive learning sample complexity in that case. It may be possible to reduce or remove this factor. On a related note, Hegedüs [3] introduces a modified MembHalving algorithm, which makes queries in a particular greedy order. By doing so, the bound decreases to $2\frac{t_0}{\log t_0}\log_2|\mathbb{C}|$ instead of $t_0 \log_2 |\mathbb{C}|$. A similar technique might be possible here, though the effect seems more difficult to quantify. Additionally, a more careful treatment of the constants in these bounds may yield significant improvements.

The present analysis requires access to an upper bound $\eta$ on the noise rate. As mentioned, it is possible to remove this assumption by a guess-and-double procedure, using a labeled validation set of size $\Omega(1/\epsilon)$. In practice, this may not be too severe, since we often use a validation set to tune parameters or estimate the final error rate anyway. Nonetheless, it would be nice to remove this requirement without sacrificing anything in dependence on $\frac{1}{\epsilon}$. In particular, it may sometimes be possible to determine whether a classifier is near-optimal using only a few carefully chosen queries.

As a final remark, exploring the connections between the present analysis and the related approaches discussed in Section 2 could prove fruitful. Thorough study of these approaches and their interrelations seems essential for a complete understanding of the label complexity of active learning.

## References


1. M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proc. of the 23rd International Conference on Machine Learning*, 2006.
2. S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.
3. T. Hegedüs. Generalized teaching dimension and the query complexity of learning. In *Proc. of the 8th Annual Conference on Computational Learning Theory*, 1995.
4. D. Angluin. Queries revisited. *Theoretical Computer Science*, 313:175–194, 2004.
5. S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.
6. S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.
7. M. Kääriäinen. Active learning in the non-realizable case. In *Proc. of the 17th International Conference on Algorithmic Learning Theory*, 2006.
8. N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
9. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
10. A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
11. Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proc. of the 35th Annual ACM Symposium on the Theory of Computing*, pages 335–344, 2003.