RATES OF CONVERGENCE IN ACTIVE LEARNING

BY STEVE HANNEKE

Carnegie Mellon University

We study the rates of convergence in generalization error achievable by active learning under various types of label noise. Additionally, we study the general problem of model selection for active learning with a nested hierarchy of hypothesis classes, and propose an algorithm whose error rate provably converges to the best achievable error among classifiers in the hierarchy at a rate adaptive to both the complexity of the optimal classifier and the noise conditions. In particular, we state sufficient conditions for these rates to be dramatically faster than those achievable by passive learning.

1. Introduction. *Active learning* refers to a family of powerful supervised learning protocols capable of producing more accurate classifiers while using a smaller number of labeled data points than traditional (passive) learning methods. Here we study a variant known as *pool-based* active learning, in which a learning algorithm is given access to a large pool of unlabeled data (i.e., only the covariates are visible), and is allowed to sequentially request the label (response variable) of any particular data points from that pool. The objective is to learn a function that accurately predicts the labels of new points, while minimizing the number of label requests. Thus, this is a type of sequential design scenario for a function estimation problem. This contrasts with passive learning, where the labeled data are sampled at random. In comparison, by more carefully selecting which points should be labeled, active learning can often significantly decrease the total amount of effort required for data annotation. This can be particularly interesting for tasks where unlabeled data are available in abundance, but label information comes only through significant effort or cost.

Recently, there have been a series of exciting advances on the topic of active learning with arbitrary classification noise (the so-called *agnostic* PAC model [23]), resulting in several new algorithms capable of achieving improved convergence rates compared to passive learning under certain conditions. The first, proposed by Balcan, Beygelzimer, and Langford [6] was the A^2 (agnostic active) algorithm, which provably never has significantly worse rates of convergence than passive learning by empirical risk minimization. This algorithm was later analyzed in de-

1

tail in [21], where it was found that a complexity measure called the *disagreement coefficient* characterizes the worst-case convergence rates achieved by A^2 for any given hypothesis class, data distribution, and best achievable error rate in the class. The next major advance was by Dasgupta, Hsu, and Monteleoni [15], who proposed a new algorithm, and proved that it improves the dependence of the convergence rates on the disagreement coefficient compared to A^2. Both algorithms are defined below in Section 3. While all of these advances are encouraging, they are limited in two ways. First, the convergence rates that have been proven for these algorithms typically only improve the dependence on the magnitude of the noise (more precisely, the noise rate of the hypothesis class), compared to passive learning. Thus, in an asymptotic sense, for nonzero noise rates these results represent at best a constant factor improvement over passive learning. Second, these results are limited to learning with a fixed hypothesis class of limited expressiveness, so that convergence to the Bayes error rate is not always a possibility.

On the first of these limitations, recent work by Castro and Nowak [13] on learning threshold classifiers discovered that if certain parameters of the noise distribution are *known* (namely, parameters related to Tsybakov's margin conditions), then we can achieve strict improvements in the asymptotic convergence rate via a specific active learning algorithm designed to take advantage of that knowledge for thresholds. Subsequently, Balcan, Broder, and Zhang [7] proved a similar result for linear separators in higher dimensions, and Castro and Nowak [13] showed related improvements for the space of boundary fragment classes (under a somewhat stronger assumption than Tsybakov's). However, these works left open the question of whether such improvements could be achieved by an algorithm that does not explicitly depend on the noise conditions (i.e., in the *agnostic* setting), and whether this type of improvement is achievable for more general families of hypothesis classes, under the usual complexity restrictions (e.g., VC class, entropy conditions, etc.). In a personal communication, John Langford and Rui Castro claimed A^2 achieves these improvements for the special case of threshold classifiers (a special case of this also appeared in [9]). However, there remained an open question of whether such rate improvements could be generalized to hold for arbitrary hypothesis classes. In Section 4, we provide this generalization. We analyze the rates achieved by A^2 under Tsybakov's noise conditions [27, 29]; in particular, we find that these rates are strictly superior to the known rates for passive learning, when the disagreement coefficient is finite. We also study a novel modification of the algorithm of Dasgupta, Hsu, and Monteleoni [15], proving that it improves upon the rates of A^2 in its dependence on the disagreement coefficient.

Additionally, in Section 5, we address the second limitation by proposing a general model selection procedure for active learning with an arbitrary structure of nested hypothesis classes. If the classes have restricted expressiveness (e.g., VC

classes), the error rate for this algorithm converges to the best achievable error by any classifier in the structure, at a rate that adapts to the noise conditions and complexity of the optimal classifier. In general, if the structure is constructed to include arbitrarily good approximations to any classifier, the error converges to the Bayes error rate in the limit. In particular, if the Bayes optimal classifier is in some class within the structure, the algorithm performs nearly as well as running an agnostic active learning algorithm on that single hypothesis class, thus preserving the convergence rate improvements achievable for that class.

2. Definitions and Notation. In the active learning setting, there is an *instance space* $\mathcal{X}$, a *label space* $\mathcal{Y} = \{-1, +1\}$, and some fixed distribution $\mathcal{D}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, with marginal $\mathcal{D}_X$ over $\mathcal{X}$. The restriction to binary classification ($\mathcal{Y} = \{-1, +1\}$) is intended to simplify the discussion; however, everything below generalizes quite naturally to multiclass classification (where $\mathcal{Y} = \{1, 2, \ldots, k\}$).

There are two sequences of random variables: $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$, where each (X_i, Y_i) pair is independent of the others, and has joint distribution $\mathcal{D}_{XY}$. However, the learning algorithm is only permitted direct access to the X_i values (unlabeled data points), and must request the Y_i values one at a time, sequentially. That is, the algorithm picks some index i to observe the Y_i value, then after observing it, picks another index i' to observe the $Y_{i'}$ label value, etc. We are interested in studying the rate of convergence of the error rate of the classifier output by the learning algorithm, in terms of the number of label requests it has made. To simplify the discussion, we will think of the data sequence as being essentially inexhaustible, and will study $(1 - \delta)$-confidence bounds on the error rate of the classifier produced by an algorithm permitted to make at most n label requests, for a fixed value $\delta \in (0, 1/2)$. The actual number of (unlabeled) data points the algorithm uses will be made clear in the proofs (typically close to the number of points needed by passive learning to achieve the stated error guarantee).

A *hypothesis class* $\mathbb{C}$ is any set of measurable classifiers $h : \mathcal{X} \to \mathcal{Y}$. We will denote by d the VC dimension of $\mathbb{C}$ [see e.g., 12, 16, 31–33]. For any measurable $h : \mathcal{X} \to \mathcal{Y}$ and distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, define the *error rate* of h as $er_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y) \sim \mathcal{D}}\{h(X) \neq Y\}$; when $\mathcal{D} = \mathcal{D}_{XY}$, we abbreviate this as $er(h)$. This simply represents the risk under the 0-1 loss. We also define the *conditional error rate*, given a set $R \subseteq \mathcal{X}$, as $er(h|R) = \mathbb{P}\{h(X) \neq Y | X \in R\}$. Let $\nu = \inf_{h \in \mathbb{C}} er(h)$, called the *noise rate* of $\mathbb{C}$. For any $x \in \mathcal{X}$, let $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$, let $h^*(x) = 2\mathbb{1}[\eta(x) \geq 1/2] - 1$, and let $\nu^* = er(h^*)$. We call h^* the *Bayes optimal classifier*, and ν^* the *Bayes error rate*. Additionally, define the *diameter* of any set of classifiers V as $diam(V) = \sup_{h_1, h_2 \in V} \mathbb{P}\{h_1(X) \neq h_2(X)\}$, and for any $\epsilon > 0$, define the diameter of the ϵ-*minimal set* of V as $diam(\epsilon; V) = diam(\{h \in V : er(h) - \inf_{h' \in V} er(h') \leq \epsilon\})$.

For a classifier h, and a sequence $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, let $er_S(h) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[h(x) \neq y]$ denote the *empirical error rate* on S, (and define $er_{\{\}}(h) = 0$ by convention). It will often be convenient to make use of sets of (index, label) pairs, where the index is used to uniquely refer to an element of the $\{X_i\}$ sequence (while conveniently also keeping track of relative ordering information); in such contexts, we will overload notation as follows. For a classifier h, and a finite set of (index, label) pairs $S = \{(i_1, y_1), (i_2, y_2), \ldots, (i_m, y_m)\} \subset \mathbb{N} \times \mathcal{Y}$, let $er_S(h) = \frac{1}{|S|} \sum_{(i,y) \in S} \mathbb{1}[h(X_i) \neq y]$, (and $er_{\{\}}(h) = 0$, as before). Thus, $er_S(h) = er_{S'}(h)$, where $S' = \{(X_i, y)\}_{(i,y) \in S}$. For the indexed *true* label sequence, $\mathcal{Z}^{(m)} = \{(1, Y_1), (2, Y_2), \ldots, (m, Y_m)\}$, we abbreviate this $er_m(h) = er_{\mathcal{Z}^{(m)}}(h)$, the empirical error on the first m data points.

In addition to the independent interest of understanding the rates achievable here, another primary interest in this setting is to quantify the achievable *improvements*, compared to *passive learning*. In this context, a passive learning algorithm can be formally defined as a function mapping the sequence $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ to a classifier $\hat{h}_n$; for instance, perhaps the most widely studied family of passive learning methods is that of *empirical risk minimization* [e.g., 24, 28, 31, 32], which return a classifier $\hat{h}_n \in \operatorname{argmin}_{h \in \mathbb{C}} er_n(h)$. For the purpose of this comparison, we review known results on passive learning in several contexts below.

2.1. *Tsybakov's Noise Conditions.* Here we describe a particular parametrization of noise distributions, relative to a hypothesis class, often referred to as Tsybakov's noise conditions [27, 29], or margin conditions. These noise conditions have recently received substantial attention in the passive learning literature, as they describe situations in which the asymptotic minimax convergence rate of passive learning is faster than the worst case $n^{-1/2}$ rate [e.g., 24, 27–29].

CONDITION 1.   *There exist finite constants $\mu > 0$ and $\kappa \geq 1$, s.t. $\forall \epsilon > 0$, $diam(\epsilon; \mathbb{C}) \leq \mu \epsilon^{\frac{1}{\kappa}}$.*   ◇

This condition is satisfied when, for example,

$$\exists \mu' > 0, \kappa \geq 1 \text{ s.t. } \exists h \in \mathbb{C} : \forall h' \in \mathbb{C}, er(h') - \nu \geq \mu' \mathbb{P}\{h(X) \neq h'(X)\}^\kappa$$

[24]. It is also satisfied when the Bayes optimal classifier is in $\mathbb{C}$ and

$$\exists \mu'' > 0, \alpha \in (0, \infty) \text{ s.t. } \forall \epsilon > 0, \mathbb{P}\{|\eta(X) - 1/2| \leq \epsilon\} \leq \mu'' \epsilon^\alpha,$$

where κ and μ are functions of α and μ'' [27, 29]; in particular, $\kappa = (1 + \alpha)/\alpha$. As we will see, the case where $\kappa = 1$ is particularly interesting; for instance,

this is the case when $h^* \in \mathbb{C}$ and $\mathbb{P}\{|\eta(X) - 1/2| > c\} = 1$ for some constant $c \in (0, 1/2)$. Informally, in many cases Condition 1 can be realized in terms of the relation between magnitude of noise and distance to the optimal decision boundary; that is, since in practice the amount of noise in a data point's label is often inversely related to the distance from the decision boundary, a small κ value may often result from having low density near the decision boundary (i.e., large margin); when this is not the case, the value of κ is often determined by how quickly $\eta(x)$ changes as x approaches the decision boundary. See [7, 13, 24, 27–29] for further interpretations of this condition.

It is known that when this condition is satisfied for some $\kappa \geq 1$ and $\mu > 0$, the passive learning method of empirical risk minimization achieves a convergence rate guarantee, holding with probability $\geq 1 - \delta$, of

$$er(\underset{h \in \mathbb{C}}{\operatorname{argmin}} \, er_n(h)) - \nu \leq c \left(\frac{d \log n + \log(1/\delta)}{n} \right)^{\frac{\kappa}{2\kappa-1}},$$

where c is a (κ and μ -dependent) constant (this follows from [24, 28]; see Appendix B, especially (17) and Lemma 5, for the details). Furthermore, for some hypothesis classes, this is known to be a tight bound (up to the log factor) on the minimax convergence rate, so that there is *no* passive learning algorithm for these classes for which we can guarantee a faster convergence rate, given that the guarantee depends on $\mathcal{D}_{XY}$ only through μ and κ [13, 29] (see also Appendix D).

2.2. *Disagreement Coefficient.* The disagreement coefficient, introduced in [21], is a measure of the complexity of an active learning problem, which has proven quite useful for analyzing the convergence rates of certain types of active learning algorithms: for example, the algorithms of [6, 11, 14, 15]. Informally, it quantifies how much disagreement there is among a set of classifiers relative to how close to some h they are. The following is a version of its definition, which we will use extensively below.

For any hypothesis class $\mathbb{C}$ and $V \subseteq \mathbb{C}$, let

$$DIS(V) = \{x \in \mathcal{X} : \exists h_1, h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\}.$$

For $r \in [0, 1]$ and measurable $h : \mathcal{X} \to \mathcal{Y}$, let

$$B(h, r) = \{h' \in \mathbb{C} : \mathbb{P}\{h(X) \neq h'(X)\} \leq r\}.$$

DEFINITION 1. *The* disagreement coefficient *of h with respect to $\mathbb{C}$ under $\mathcal{D}_X$ is defined as*

$$\theta_h = \sup_{r > r_0} \frac{\mathbb{P}(DIS(B(h, r)))}{r},$$

where $r_0 = 0$ (though see Appendix A.1 for alternative possibilities for r_0).   ⋄

DEFINITION 2. *We further define the disagreement coefficient for the hypothesis class $\mathbb{C}$ with respect to the target distribution $\mathcal{D}_{XY}$ as $\theta = \liminf_{k \to \infty} \theta_{h^{[k]}}$, where $\{h^{[k]}\}$ is any sequence in $\mathbb{C}$ with $er(h^{[k]})$ monotonically decreasing to ν; (by convention, take every $h^{[k]} \in \underset{h \in \mathbb{C}}{\arg\min}\, er(h)$ if the minimum is achieved).*   ◇

In Definition 1, it is conceivable that $DIS(B(h, r))$ may sometimes not be measurable. In such cases, we can define $\mathbb{P}(DIS(B(h, r)))$ as the *outer* measure [30], so that it remains well defined. We continue this practice below, letting $\mathbb{P}$ and $\mathbb{E}$ (and indeed any reference to "probability") refer to the outer expectation and measure in any context for which this is necessary.

Because of its simple intuitive interpretation, measuring the amount of disagreement in a local neighborhood of some classifier h, the disagreement coefficient has the wonderful property of being relatively simple to calculate for a wide range of learning problems, especially when those problems have a natural geometric representation. To illustrate this, we will go through a few simple examples from [21].

Consider the hypothesis class of thresholds h_z on the interval $[0, 1]$ (for $z \in (0, 1)$), where $h_z(x) = +1$ iff $x \geq z$. Furthermore, suppose $\mathcal{D}_X$ is uniform on $[0, 1]$. In this case, it is clear that the disagreement coefficient is 2, since for sufficiently small r, the region of disagreement of $B(h_z, r)$ is $[z - r, z + r)$, which has probability mass $2r$. In other words, since the disagreement region grows with r in two disjoint directions, each at rate 1, we have $\theta_{h_z} = 2$.

As a second example, consider the disagreement coefficient for *intervals* on $[0, 1]$. As before, let $\mathcal{X} = [0, 1]$ and $\mathcal{D}_X$ be uniform, but this time $\mathbb{C}$ is the set of intervals $h_{[a,b]}$ such that for $x \in [0, 1]$, $h_{[a,b]}(x) = +1$ iff $x \in [a, b]$ (for $0 < a < b < 1$). In contrast to thresholds, the disagreement coefficients $\theta_{h_{[a,b]}}$ for the space of intervals vary widely depending on the particular $h_{[a,b]}$. Specifically, we have $\theta_{h_{[a,b]}} = \max\left\{\frac{1}{b-a}, 4\right\}$. To see this, note that when $0 < r < b - a$, every interval in $B(h_{[a,b]}, r)$ has its lower and upper boundaries within r of a and b, respectively; thus, $\mathbb{P}(DIS(B(h_{[a,b]}, r))) \leq 4r$, with equality for sufficiently small r. However, when $r > b - a$, *every* interval of width $\leq r - (b-a)$ is in $B(h_{[a,b]}, r)$, so that $\mathbb{P}(DIS(B(h_{[a,b]}, r))) = 1$.

As a slightly more involved example, [21] studies the scenario where $\mathcal{X}$ is the surface of the origin-centered unit sphere in $\mathbb{R}^d$ for $d > 2$, $\mathbb{C}$ is the space of all linear separators whose decision surface passes through the origin, and $\mathcal{D}_X$ is the uniform distribution on $\mathcal{X}$; in this case, it turns out $\forall h \in \mathbb{C}$ the disagreement coefficient θ_h satisfies

$$\frac{\pi}{4}\sqrt{d} \leq \theta_h \leq \pi\sqrt{d}.$$

The disagreement coefficient has many interesting properties that can help to bound its value for a given hypothesis class and distribution. We list a few elemen-

tary properties below. Their proofs, which are quite short and follow directly from the definition, are left as easy exercises.

LEMMA 1. *[Close Marginals][21] Suppose $\exists \lambda \in (0,1]$ s.t. for any measurable set $A \subseteq \mathcal{X}$, $\lambda \mathbb{P}_{\mathcal{D}_X}(A) \leq \mathbb{P}_{\mathcal{D}'_X}(A) \leq \frac{1}{\lambda} \mathbb{P}_{\mathcal{D}_X}(A)$. Let $h : \mathcal{X} \to \mathcal{Y}$ be a measurable classifier, and suppose θ_h and θ'_h are the disagreement coefficients for h with respect to $\mathbb{C}$ under $\mathcal{D}_X$ and $\mathcal{D}'_X$ respectively. Then*

$$\lambda^2 \theta_h \leq \theta'_h \leq \frac{1}{\lambda^2} \theta_h. \qquad \diamond$$

LEMMA 2. *[Finite Mixtures] Suppose $\exists \alpha \in [0,1]$ s.t. for any measurable set $A \subseteq \mathcal{X}$, $\mathbb{P}_{\mathcal{D}_X}(A) = \alpha \mathbb{P}_{\mathcal{D}_1}(A) + (1-\alpha)\mathbb{P}_{\mathcal{D}_2}(A)$. For a measurable $h : \mathcal{X} \to \mathcal{Y}$, let $\theta_h^{(1)}$ be the disagreement coefficient with respect to $\mathbb{C}$ under $\mathcal{D}_1$, $\theta_h^{(2)}$ be the disagreement coefficient with respect to $\mathbb{C}$ under $\mathcal{D}_2$, and θ_h be the disagreement coefficient with respect to $\mathbb{C}$ under $\mathcal{D}_X$. Then*

$$\theta_h \leq \theta_h^{(1)} + \theta_h^{(2)}. \qquad \diamond$$

LEMMA 3. *[Finite Unions] Suppose $h \in \mathbb{C}_1 \cap \mathbb{C}_2$ is a classifier s.t. the disagreement coefficient with respect to $\mathbb{C}_1$ under $\mathcal{D}_X$ is $\theta_h^{(1)}$ and with respect to $\mathbb{C}_2$ under $\mathcal{D}_X$ is $\theta_h^{(2)}$. Then if θ_h is the disagreement coefficient with respect to $\mathbb{C} = \mathbb{C}_1 \cup \mathbb{C}_2$ under $\mathcal{D}_X$, we have that*

$$\max\left\{\theta_h^{(1)}, \theta_h^{(2)}\right\} \leq \theta_h \leq \theta_h^{(1)} + \theta_h^{(2)}.$$

In fact, even if $h \notin \mathbb{C}_1 \cap \mathbb{C}_2$, we still have $\theta_h \leq \theta_h^{(1)} + \theta_h^{(2)} + 2$. $\qquad \diamond$

See [8, 11, 15, 17, 21, 35] for further discussions of various uses of the disagreement coefficient and related notions and extensions in active learning. In particular, Friedman [17] proves that any hypothesis class and distribution satisfying certain general regularity conditions will admit finite constant bounds on θ. Also, Wang [35] bounds the disagreement coefficient for certain nonparametric hypothesis classes, characterized by smoothness of their decision surfaces. Additionally, Beygelzimer, Dasgupta, and Langford [11] present an interesting analysis using a natural extension of the disagreement coefficient to study active learning with a larger family of loss functions beyond 0-1 loss.

The disagreement coefficient has deep connections to several other quantities, such as doubling dimension [26] and VC dimension [31]. Additionally, a related quantity, referred to as the "capacity function," was studied in the 80s by Alexander in the passive learning literature, in the context of ratio-type empirical processes

[2–4], and recently was further developed by Giné and Koltchinskii [18]; interestingly, in this latter work, Giné and Koltchinskii study a localized version of the capacity function, which in our present context can essentially be viewed as the function $\tau(r) = \mathbb{P}(DIS(B(h,r)))/r$, so that $\theta_h = \sup_{r > r_0} \tau(r)$.

3. General Algorithms. We begin the discussion of the algorithms we will analyze by noting the underlying inspiration that unifies them. Specifically, at this writing, all of the published general-purpose agnostic active learning algorithms achieving nontrivial improvements are derivatives of a basic technique proposed by Cohn, Atlas, and Ladner [14] for the realizable active learning problem. Under the assumption that there exists a perfect classifier in $\mathbb{C}$, they proposed an algorithm which processes unlabeled data points in sequence, and for each one it determines whether there is a classifier in $\mathbb{C}$ consistent with all previously observed labels that predicts $+1$ for this new point *and* one that predicts -1 for this new point; if so, the algorithm requests the label, and otherwise it does not request the label; after n label requests, the algorithm returns any classifier consistent with all observed labels. In some sense, this algorithm corresponds to the very least we could expect of an active learning algorithm, as it never requests the label of a point it can derive from known information, but otherwise makes no effort to search for informative data points. The idea is appealing, not only for its simplicity, but also for its extremely efficient use of unlabeled data; in fact, under the stated assumption, the algorithm produces a classifier consistent with the labels of *all* of the unlabeled data it processes, including those it does *not* request the labels of.

We can equivalently think of this algorithm as maintaining two sets: $V \subseteq \mathbb{C}$ is the set of candidate hypotheses still under consideration, and $R = DIS(V)$ is their region of disagreement. We can then think of the algorithm as requesting a random labeled point from the conditional distribution of $\mathcal{D}_{XY}$ given that $X \in R$, and subsequently removing from V any classifier inconsistent with the observed label. A formal definition of the algorithm is given as follows.

Algorithm 0

Input: hypothesis class $\mathbb{C}$, label budget n
Output: classifier $\hat{h}_n \in \mathbb{C}$

---

0. $V_0 \leftarrow \mathbb{C}$, $t \leftarrow 0$
1. For $m = 1, 2, \ldots$
2.   If $X_m \in DIS(V_t)$,
3.   Request Y_m
4. $t \leftarrow t + 1$
5. $V_t \leftarrow \{h \in V_{t-1} : h(X_m) = Y_m\}$
6.   If $t = n$ or $\{m' > m : X_{m'} \in DIS(V_t)\} = \emptyset$, Return any $\hat{h}_n \in V_t$

The algorithms described below for the problem of active learning with label noise each represent noise-robust variants of this basic idea. They work to reduce the set of candidate hypotheses, while only requesting the labels of points in the region of disagreement of these candidates. The trick is to only remove a classifier from the candidate set once we have high statistical confidence that it is worse than some other candidate classifier so that we never remove the best classifier. However, the two algorithms differ somewhat in the details of how that confidence is calculated.

3.1. *Algorithm 1.* The first noise-robust algorithm we study, originally proposed by Balcan, Beygelzimer, and Langford [6], is typically referred to as A^2 for *Agnostic Active*. This was historically the first general-purpose agnostic active learning algorithm shown to achieve improved error guarantees for certain learning problems in certain ranges of n and ν. Below is a variant of this algorithm. It is defined in terms of two functions: UB and LB. These represent upper and lower confidence bounds on the error rate of a classifier from $\mathbb{C}$ with respect to an arbitrary sampling distribution, as a function of a labeled sequence sampled according to that distribution. Some steps in the algorithm require calculating certain probabilities, such as $\mathbb{P}(DIS(V))$ or $\mathbb{P}(R)$; later, we discuss replacing these with appropriate estimators.

Algorithm 1
Input: hypothesis class $\mathbb{C}$, label budget n, confidence δ, functions UB and LB
Output: classifier $\hat{h}_n$

---

0. $V \leftarrow \mathbb{C}, R \leftarrow DIS(\mathbb{C}), Q \leftarrow \emptyset, m \leftarrow 0$
1. For $t = 1, 2, \ldots, n$
2.   If $\mathbb{P}(DIS(V)) \leq \frac{1}{2}\mathbb{P}(R)$
3. $R \leftarrow DIS(V); Q \leftarrow \emptyset$
4.   If $\mathbb{P}(R) \leq 2^{-n}$, Return any $\hat{h}_n \in V$
5. $m \leftarrow \min\{m' > m : X_{m'} \in R\}$
6.   Request Y_m and let $Q \leftarrow Q \cup \{(m, Y_m)\}$
7. $V \leftarrow \{h \in V : LB(h, Q, \delta/n) \leq \min_{h' \in V} UB(h', Q, \delta/n)\}$
8. $h_t \leftarrow \operatorname*{argmin}_{h \in V} UB(h, Q, \delta/n)$
9. $\beta_t \leftarrow (UB(h_t, Q, \delta/n) - \min_{h \in V} LB(h, Q, \delta/n))\mathbb{P}(R)$
10. Return $\hat{h}_n = h_{\hat{t}}$, where $\hat{t} = \operatorname*{argmin}_{t \in \{1,2,\ldots,n\}} \beta_t$

The intuitive motivation behind the algorithm is the following. It focuses on reducing the set of candidate hypotheses V, while being careful not to throw away the best classifier $h_{\mathbb{C}}^* = \operatorname{argmin}_{h \in \mathbb{C}} er(h)$ (supposing, for this informal expla-

nation, that $h_{\mathbb{C}}^*$ exists). Given that this is satisfied at any given time in the algo-rithm, it makes sense to focus our samples to the region $DIS(V)$, since a clas-sifier $h_1 \in V$ has smaller error rate than another classifier $h_2 \in V$ if and only if it has smaller conditional error rate given $DIS(V)$. For this reason, on each round, we seek to remove from V any h for which our confidence bounds indi-cate that $er(h|DIS(V)) > er(h_{\mathbb{C}}^*|DIS(V))$. However, so that we can make use of known results for i.i.d. samples, we freeze the sampling region $R \supseteq DIS(V)$ and collect an i.i.d. sample from the conditional given this region, updating the region only when doing so allows us to further significantly focus the samples; for this same reason, we also reset the collection of samples Q every time we update the region R, so that it represents samples from the conditional given R. Finally, we maintain the values β_t, which represent confidence upper bounds on $er(h_t) - \nu = (er(h_t|R) - er(h_{\mathbb{C}}^*|R))\mathbb{P}(R)$, and we return the h_t minimizing this confidence bound; note that it does not suffice to return h_n, since the final Q set might be small.

As long as the confidence bounds UB and LB satisfy (overloading notation in the natural way)

$$\mathbb{P}_{Z \sim \mathcal{D}^m}\{\forall h \in \mathbb{C}, LB(h, Z, \delta') \leq er_{\mathcal{D}}(h) \leq UB(h, Z, \delta')\} \geq 1 - \delta'$$

for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and any $\delta' \in (0, 1)$, and UB and LB converge to each other as m grows, it is known that a $1 - \delta$ confidence bound on $er(\hat{h}_n) - \nu$ converges to 0 [6]. For instance, Balcan, Beygelzimer, and Langford [6] suggest defining these functions based on classic results on uniform convergence rates in passive learning [31], such as

$$(1) \qquad UB(h, Q, \delta') = \min\{er_Q(h) + G(|Q|, \delta'), 1\},$$
$$LB(h, Q, \delta') = \max\{er_Q(h) - G(|Q|, \delta'), 0\},$$

where $G(m, \delta') = \frac{1}{m} + \sqrt{\frac{\ln \frac{4}{\delta'} + d \ln \frac{2em}{d}}{m}}$ for $m \geq d$, and by convention $G(m, \delta') = \infty$ for $m < d$. This choice of UB and LB is motivated by the following lemma, due to Vapnik [32].

LEMMA 4. *For any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and any $\delta' \in (0, 1)$ and $m \in \mathbb{N}$, with probability $\geq 1 - \delta'$ over the draw of $Z \sim \mathcal{D}^m$, every $h \in \mathbb{C}$ satisfies*

$$(2) \qquad |er_Z(h) - er_{\mathcal{D}}(h)| \leq G(m, \delta').$$

$\diamond$

To avoid computational issues, instead of explicitly representing the sets V and R, we may implicitly represent them as a set of constraints imposed by the condi-tion in Step 7 of previous iterations. We may also replace $\mathbb{P}(DIS(V))$ and $\mathbb{P}(R)$

by estimates, since these quantities can be estimated to arbitrary precision with arbitrarily high confidence using only *unlabeled* data. Specifically, the convergence rates proven below can be preserved up to constant factors by replacing these quantities with confidence bounds depending on a finite number of unlabeled data points; for completeness, the details are given in Appendix C. As for the number of unlabeled data points required by the above algorithm itself, note that if $\mathbb{P}(DIS(V))$ becomes small, it will use a large number of unlabeled data points; however, $\mathbb{P}(DIS(V))$ being small also indicates $er(\hat{h}_n) - \nu$ is small (and indeed β_t). In particular, to get an excess error rate of ϵ, the algorithm will generally require a number of unlabeled data points only polynomial in $1/\epsilon$; also, the condition in Step 4 guarantees the total number of unlabeled data points used by the algorithm is bounded with high probability. For comparison, recall that passive learning typically requires a number of *labeled* data points polynomial in $1/\epsilon$.

3.2. *Algorithm 2.*   The second noise-robust algorithm we study was originally proposed by Dasgupta, Hsu, and Monteleoni [15]. It uses a type of constrained passive learning subroutine, LEARN, defined as follows for two sets of labeled data points, $\mathcal{L}$ and Q.

$$\text{LEARN}_{\mathbb{C}}(\mathcal{L}, Q) = \underset{h \in \mathbb{C}: er_{\mathcal{L}}(h) = 0}{\text{argmin}} \; er_Q(h).$$

By convention, if no $h \in \mathbb{C}$ has $er_{\mathcal{L}}(h) = 0$, $\text{LEARN}_{\mathbb{C}}(\mathcal{L}, Q) = \varnothing$. The algorithm is formally defined below, in terms of a sequence of estimators Δ_m, defined later.

Algorithm 2
Input: hypothesis class $\mathbb{C}$, label budget n, confidence δ, functions Δ_m
Output: classifier $\hat{h}_n$, sets of (index, label) pairs $\mathcal{L}$ and Q

---

0. $\mathcal{L} \leftarrow \emptyset, Q \leftarrow \emptyset$
1. For $m = 1, 2, \ldots$
2.   If $|Q| = n$ or $m > 2^n$, Return $\hat{h}_n = \text{LEARN}_{\mathbb{C}}(\mathcal{L}, Q)$ along with $\mathcal{L}$ and Q
3.   For each $y \in \{-1, +1\}$, let $h^{(y)} = \text{LEARN}_{\mathbb{C}}(\mathcal{L} \cup \{(m, y)\}, Q)$
4.   If some y has $h^{(-y)} = \varnothing$ or
          $er_{\mathcal{L} \cup Q}(h^{(-y)}) - er_{\mathcal{L} \cup Q}(h^{(y)}) > \Delta_{m-1}(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta)$
5.   Then $\mathcal{L} \leftarrow \mathcal{L} \cup \{(m, y)\}$
6.   Else Request the label Y_m and let $Q \leftarrow Q \cup \{(m, Y_m)\}$

The algorithm maintains two sets of labeled data points: $\mathcal{L}$ and Q. The set Q represents points of which we have requested the labels. The set $\mathcal{L}$ represents the remaining points, and the labels of points in $\mathcal{L}$ are *inferred*. Specifically, suppose (inductively) that at some time m we have that every $(i, y) \in \mathcal{L}$ has $h_{\mathbb{C}}^*(X_i) = y$, where $h_{\mathbb{C}}^* = \text{argmin}_{h \in \mathbb{C}} er(h)$ (supposing the min is achieved, for this informal

motivation). At any point, we can be fairly confident that $h_{\mathbb{C}}^*$ will have relatively small empirical error rate. Thus, if all of the classifiers h with $er_{\mathcal{L}}(h) = 0$ and $h(X_m) = -y$ have relatively large empirical error rates compared to some h with $er_{\mathcal{L}}(h) = 0$ and $h(X_m) = y$, we can confidently infer that $h_{\mathbb{C}}^*(X_m) = y$. Note that this is not the *true* label Y_m, but a sort of "denoised" version of it. Once we infer this label, since we are already confident that this is the $h_{\mathbb{C}}^*$ label, and $h_{\mathbb{C}}^*$ is the classifier we wish to compete with, we simply add this label as a *constraint*: that is, we require every classifier under consideration in the future to have $h(X_m) = h_{\mathbb{C}}^*(X_m)$. This is how elements of $\mathcal{L}$ are added. On the other hand, if we cannot confidently infer $h_{\mathbb{C}}^*(X_m)$, because some classifiers labeling opposite this also have relatively small empirical error rates, then we simply request the label Y_m and add it to the set Q. Note that in order to make this comparison, we needed to be able to calculate the differences of empirical error rates; however, as long as we only consider the set of classifiers h that *agree* on the labels in $\mathcal{L}$, we will have $er_{\mathcal{L} \cup Q}(h_1) - er_{\mathcal{L} \cup Q}(h_2) = er_m(h_1) - er_m(h_2)$, for any two such classifiers h_1 and h_2, where $m = |\mathcal{L} \cup Q|$.

The key to the above argument is carefully choosing a threshold for how large the difference in empirical error rates needs to be before we can confidently infer the label. For this purpose, Algorithm 2 is defined in terms of a function, $\Delta_m(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta)$, representing a threshold for a type of hypothesis test. This threshold must be set carefully, since the sequence of labeled data points corresponding to $\mathcal{L} \cup Q$ is not actually an i.i.d. sample from $\mathcal{D}_{XY}$. Dasgupta, Hsu, and Monteleoni [15] suggest defining this function as

$$(3) \quad \Delta_m(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta) = \beta_m^2 + \beta_m \left(\sqrt{er_{\mathcal{L} \cup Q}(h^{(y)})} + \sqrt{er_{\mathcal{L} \cup Q}(h^{(-y)})} \right),$$

where $\beta_m = \sqrt{\frac{4 \ln(8m(m+1)\mathcal{S}(\mathbb{C}, 2m)^2/\delta)}{m}}$ and $\mathcal{S}(\mathbb{C}, 2m)$ is the shatter coefficient [e.g., 16, 32]; this suggestion is based on a confidence bound they derive, and they prove the correctness of the algorithm with this definition, meaning that the $1 - \delta$ confidence bound on its error rate converges to ν as $n \to \infty$. For now we will focus on the first return value (the classifier), leaving the others for Section 5, where they will be useful for chaining multiple executions together.

4. Convergence Rates. In both of the above cases, one can prove guarantees stating that neither algorithm's convergence rates are ever significantly worse than passive learning by empirical risk minimization [6, 15]. However, it is even more interesting to discuss situations in which one can prove error rate guarantees for these algorithms significantly *better* than those achievable by passive learning. In this section, we begin by reviewing known results on these potential improvements, stated in terms of the disagreement coefficient; we then proceed to discuss new

results for Algorithm 1 and a novel variant of Algorithm 2, and describe the convergence rates achieved by these methods in terms of the disagreement coefficient and Tsybakov's noise conditions.

To simplify the presentation, for the remainder of this paper we will restrict the discussion to situations with $\theta > 0$ (and therefore $\mathbb{C}$ with $d > 0$ too). Handling the extra case of $\theta = 0$ is a trivial matter, since $\theta = 0$ would imply that any proper learning algorithm achieves excess error 0 for all values of n.

4.1. *The Disagreement Coefficient and Active Learning: Basic Results.* Before going into the results for general distributions $\mathcal{D}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, it will be instructive to first look at the special case when the noise rate is zero. Understanding how the disagreement coefficient enters into the analysis of this simpler case may aid in digestion of the theorems and proofs for the general case presented later, where it plays an essentially analogous role. Most of the major ingredients of the proofs for the general case can be found in this special case, albeit in a much simpler form. Although this result has not previously been published, the proof is essentially analogous to (one case of) the analysis of Algorithm 1 in [21].

THEOREM 1. *Let $f \in \mathbb{C}$ be such that $er(f) = 0$ and $\theta_f < \infty$. $\forall n \in \mathbb{N}$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over the draw of the unlabeled data, the classifier $\hat{h}_n$ returned by Algorithm 0 after n label requests satisfies*

$$er(\hat{h}_n) \leq 2 \cdot exp \left\{ -\frac{n}{12\theta_f \left(d \ln \left(22\theta_f \right) + \ln \left(3n/\delta \right) \right)} \right\}.$$

$\diamond$

PROOF OF THEOREM 1. As in the algorithm, let V_t denote the set of classifiers in $\mathbb{C}$ consistent with the first t label requests. If $\mathbb{P}(DIS(V_t)) > 0$ for all values of t in the algorithm, then with probability 1 the algorithm uses all n label requests. Technically, each claim below should be followed by the phrase, "unless $\mathbb{P}(DIS(V_t)) = 0$ for some $t \leq n$, in which case $er(\hat{h}_n) = 0$ so the bound trivially holds." However, to simplify the presentation, we will make this special case implicit, and will not mention it further.

The high-level outline of this proof is to use $\mathbb{P}(DIS(V_t))$ as an upper bound on $\sup_{h \in V_t} er(h)$, and then show $\mathbb{P}(DIS(V_t))$ is halved roughly every $\lambda = \tilde{O}(\theta_f d)$ label requests. Thus, after roughly $\tilde{O}(\theta_f d \log(1/\epsilon))$ label requests, any $h \in V_t$ should have $er(h) \leq \epsilon$.

Specifically, let $\lambda_n = \lceil 8\theta_f (d \ln(8e\theta_f) + \ln(2n/\delta)) \rceil$. If $n \leq \lambda_n$, the bound in the theorem statement trivially holds, since the right side exceeds 1; otherwise, consider some non-negative $t \leq n - \lambda_n$ and $t' = t + \lambda_n$. Let X_{m_t} denote the point

corresponding to the tth label request, and let $X_{m_{t'}}$ denote the point corresponding to label request number t'. It must be that

$$|\{X_{m_t+1}, X_{m_t+2}, \ldots, X_{m_{t'}}\} \cap DIS(V_t)| \geq \lambda_n,$$

which means there is an i.i.d. sample of size λ_n, with distribution equivalent to the conditional of X given $\{X \in DIS(V_t)\}$, contained in $\{X_{m_t+1}, \ldots, X_{m_{t'}}\}$: namely, the first λ_n points in this subsequence that are in $DIS(V_t)$.

Now recall that, by classic results from the passive learning literature [e.g., 5], this implies that on an event $E_{\delta,t}$ holding with probability $1 - \delta/n$,

$$\sup_{h \in V_{t'}} er(h|DIS(V_t)) \leq 2 \frac{d \ln \frac{2e\lambda_n}{d} + \ln \frac{2n}{\delta}}{\lambda_n}.$$

Also note that λ_n was defined (with express purpose) so that

$$2 \frac{d \ln \frac{2e\lambda_n}{d} + \ln \frac{2n}{\delta}}{\lambda_n} \leq 1/(2\theta_f).$$

Recall that, since $er(f) = 0$, we have $er(h) = \mathbb{P}(h(X) \neq f(X))$. Since $f \in V_{t'} \subseteq V_t$, this means for any $h \in V_{t'}$ we have $\{x : h(x) \neq f(x)\} \subseteq DIS(V_t)$, and thus

$$\sup_{h \in V_{t'}} \mathbb{P}(h(X) \neq f(X)) = \sup_{h \in V_{t'}} \mathbb{P}(h(X) \neq f(X)|X \in DIS(V_t))\mathbb{P}(DIS(V_t))$$

$$= \sup_{h \in V_{t'}} er(h|DIS(V_t))\mathbb{P}(DIS(V_t)) \leq \mathbb{P}(DIS(V_t))/(2\theta_f).$$

So $V_{t'} \subseteq B(f, \mathbb{P}(DIS(V_t))/(2\theta_f))$, and therefore by monotonicity of $\mathbb{P}(DIS(\cdot))$ and the definition of θ_f

$$\mathbb{P}(DIS(V_{t'})) \leq \mathbb{P}\left(DIS\Big(B(f, \mathbb{P}(DIS(V_t))/(2\theta_f)) \Big) \right) \leq \mathbb{P}(DIS(V_t))/2.$$

By a union bound, $E_{\delta,t}$ holds for every $t \in \{i\lambda_n : i \in \{0, 1, \ldots, \lfloor n/\lambda_n \rfloor - 1\}\}$ with probability $\geq 1 - \delta$. On these events, if $n \geq \lambda_n \lceil \log_2(1/\epsilon) \rceil$, then (by induction)

$$\sup_{h \in V_n} er(h) \leq \mathbb{P}(DIS(V_n)) \leq \epsilon.$$

Solving for ϵ in terms of n gives the result (with a slight increase in constants due to relaxing the ceiling functions). □

4.2. *Known Results on Convergence Rates for Agnostic Active Learning.* We will now describe the known results for agnostic active learning algorithms, starting with Algorithm 1. The key to the potential convergence rate improvements of Algorithm 1 is that, as the region of disagreement R decreases in measure, the error difference $er(h|R) - er(h'|R)$ of any classifiers $h, h' \in V$ under the *conditional* sampling distribution (given R) can become significantly larger (by a factor of $\mathbb{P}(R)^{-1}$) than $er(h) - er(h')$, making it significantly easier to determine which of the two is worse using a sample of labeled data. In particular, [21] developed a technique for analyzing this type of algorithm, and adapting that analysis to the above definition of Algorithm 1 results in the following guarantee.

THEOREM 2. *[21] Let $\hat{h}_n$ be the classifier returned by Algorithm 1 when allowed n label requests, using the bounds (1) and confidence parameter $\delta \in (0, 1/2)$. Then there exists a finite universal constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu \leq c\sqrt{\frac{\nu^2\theta^2(d\log n + \log\frac{1}{\delta})\log\frac{n+2\nu\theta}{\nu\theta}}{n}} + 2exp\left\{-\frac{n}{c\theta^2\left(d\log\theta + \log\frac{n}{\delta}\right)}\right\}.$$

$\diamond$

Similarly, the key to improvements from Algorithm 2 is that as the number m of processed unlabeled data points increases, we only need to request the labels of those data points in the region of disagreement of the set of classifiers with near-optimal empirical error rates. Thus, if the region of disagreement of classifiers with excess error $\leq \epsilon$ shrinks as ϵ shrinks, we expect the frequency of label requests to shrink as m increases. Since we are careful not to discard the best classifier, and the excess error rate of a classifier can be bounded in terms of the Δ_m function, we end up with a bound on the excess error which is converging in m, the number of *unlabeled* data points processed, even though we request a number of labels growing slower than m. When this situation occurs, we expect Algorithm 2 will provide an improved convergence rate compared to passive learning. Dasgupta, Hsu, and Monteleoni [15] prove the following convergence rate guarantee.

THEOREM 3. *[15] Let $\hat{h}_n$ be the classifier returned by Algorithm 2 when allowed n label requests, using the threshold (3), and confidence parameter $\delta \in (0, 1/2)$. Then there exists a finite universal constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu \leq c\sqrt{\frac{\nu^2\theta(d\log\frac{n+2\nu\theta}{\nu\theta} + \log\frac{1}{\delta})}{n}} + c\left(d + \log\frac{1}{\delta}\right)exp\left\{-\sqrt{\frac{n}{c\theta(d+\log\frac{1}{\delta})}}\right\}.$$

$\diamond$

Note that, among other changes, this bound improves the dependence on the disagreement coefficient, θ, compared to the bound for Algorithm 1. In both cases, for certain ranges of θ, ν, and n, these bounds can represent significant improvements in the excess error guarantees, compared to the corresponding guarantees possible for passive learning. However, in both cases, when $\nu > 0$ these bounds have an *asymptotic* dependence on n of $\tilde{\Theta}(n^{-1/2})$, which is no better than the convergence rates achievable by passive learning (e.g., by empirical risk minimization). Thus, there remains the question of whether either algorithm can achieve asymptotic convergence rates strictly superior to passive learning for distributions with nonzero noise rates. This is the topic we turn to next.

4.3. *Active Learning under Tsybakov's Noise Conditions.* It is known that for most nontrivial $\mathbb{C}$, for any n and $\nu > 0$, for every active learning algorithm there is some distribution with noise rate ν for which we can guarantee excess error no better than $\propto \nu n^{-1/2}$ [22]; that is, the $n^{-1/2}$ asymptotic dependence on n in the above bounds matches the corresponding minimax rate, and thus cannot be improved as long as the bounds depend on $\mathcal{D}_{XY}$ only via ν (and θ). Therefore, if we hope to discover situations in which these algorithms have strictly superior asymptotic dependence on n, we will need to allow the bounds to depend on a more detailed description of the noise distribution than simply the noise rate ν.

As previously mentioned, one way to describe a noise distribution using a more detailed parametrization is to use Tsybakov's noise conditions (Condition 1). In the context of passive learning, this allows one to describe situations in which the rate of convergence is between n^{-1} and $n^{-1/2}$, even when $\nu > 0$. This raises the natural question of how these active learning algorithms perform when the noise distribution satisfies this condition with finite μ and κ parameter values. In many ways, it seems active learning is particularly well-suited to exploit these more favorable noise conditions, since they imply that as we eliminate suboptimal classifiers, the diameter of the remaining set shrinks; thus, for finite θ values, the region of disagreement should also be shrinking, allowing us to focus the samples in a smaller region and accelerate the convergence.

Focusing on the special case of learning one-dimensional threshold classifiers under a certain uniform marginal distribution, Castro and Nowak [13] studied conditions related to Condition 1. In particular, they studied a threshold-learning algorithm that, unlike the algorithms described here, takes κ as *input*, and found its convergence rate to be $\propto \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}$ when $\kappa > 1$, and $\exp\{-cn\}$ for some (μ-dependent) constant c, when $\kappa = 1$. Note that this improves over the $n^{-\frac{\kappa}{2\kappa-1}}$ rates achievable in passive learning [13, 29]. Subsequently, Balcan, Broder, and Zhang [7] proved an analogous positive result for higher-dimensional linear separators, and Castro and Nowak [13] additionally showed a related result for bound-

ary fragment classes (see below); in both cases, the algorithm depends explicitly on the noise parameters. Later, in a personal communication, Langford and Castro claimed that in fact Algorithm 1 achieves this rate (up to log factors) for the one-dimensional thresholds problem, leading to speculation that perhaps these improvements are achievable in the general case as well (under conditions on the disagreement coefficient). Castro and Nowak [13] also prove that a value $\propto n^{-\frac{\kappa}{2\kappa-2}}$ (or $\exp\{-c'n\}$, for some c', when $\kappa = 1$) is also a *lower bound* on the minimax rate for the threshold learning problem. In fact, a similar proof to theirs can be used to show this same lower bound holds for any nontrivial $\mathbb{C}$. For completeness, a proof of this more general result is included in Appendix D.

Other than the few specific results mentioned above, it was not previously known whether Algorithm 1 or Algorithm 2, or indeed *any* active learning algorithm, generally achieves convergence rates that exhibit these types of improvements.

4.4. *Adaptive Rates in Active Learning: Algorithm 1.*   The above observations open the question of whether these algorithms, or variants thereof, improve this asymptotic dependence on n. It turns out this is indeed possible. Specifically, we have the following result for Algorithm 1.

THEOREM 4.   *Let $\hat{h}_n$ be the classifier returned by Algorithm 1 when allowed n label requests, using the bounds (1) and confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Condition 1. Then there exists a finite (κ- and μ-dependent) constant c such that, for any $n \in \mathbb{N}$, with probability $\geq 1 - \delta$,*

$$er(\hat{h}_n) - \nu \leq \begin{cases} 2 \cdot exp\left\{-\frac{n}{c\theta^2(d\log n + \log(1/\delta))}\right\}, & when \ \kappa = 1 \\ c\left(\frac{\theta^2(d\log n + \log(1/\delta))\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}, & when \ \kappa > 1 \end{cases}.$$

◇

PROOF OF THEOREM 4. We will proceed by bounding the *label complexity*, or size of the label budget n that is sufficient to guarantee, with high probability, that the excess error of the returned classifier will be at most ϵ (for arbitrary $\epsilon > 0$); with this in hand, we can simply bound the inverse of the function to get the result in terms of a bound on excess error.

Throughout this proof (and proofs of later results in this paper), we will make frequent use of basic facts about $er(h|R)$. In particular, for any classifiers h, h' and set $R \subseteq \mathcal{X}$, we have $er(h) = er(h|R)\mathbb{P}(R) + er(h|\mathcal{X} \setminus R)\mathbb{P}(\mathcal{X} \setminus R)$; also, if $\{x : h(x) \neq h'(x)\} \subseteq R$, we have $er(h|\mathcal{X} \setminus R) - er(h'|\mathcal{X} \setminus R) = 0$ and therefore $er(h) - er(h') = (er(h|R) - er(h'|R))\mathbb{P}(R)$.

Note that, by Lemma 4 and a union bound, on an event of probability $1 - \delta$, (2) holds with $\delta' = \delta/n$ for every set Q, relative to the conditional distribution given its

respective R set, for any value of n. For the remainder of this proof, we assume that this $1 - \delta$ probability event occurs. In particular, this means that for every $h \in \mathbb{C}$ and every Q set in the algorithm, $LB(h, Q, \delta/n) \leq er(h|R) \leq UB(h, Q, \delta/n)$, for the set R that Q is sampled under.

Our first task is to show that we never remove the "good" classifiers from V. We only remove a classifier h from V if $h' = \operatorname{argmin}_{h' \in V} UB(h', Q, \delta/n)$ has $LB(h, Q, \delta/n) > UB(h', Q, \delta/n)$. Each $h \in V$ has $\{x : h(x) \neq h'(x)\} \subseteq DIS(V) \subseteq R$, so that

$$UB(h', Q, \delta/n) - LB(h, Q, \delta/n) \geq er(h'|R) - er(h|R) = \frac{er(h') - er(h)}{\mathbb{P}(R)}.$$

Thus, for any $h \in V$ with $er(h) \leq er(h')$, $UB(h', Q, \delta/n) - LB(h, Q, \delta/n) \geq er(h'|R) - er(h|R) = (er(h') - er(h))/\mathbb{P}(R) \geq 0$, so that on any given round of the algorithm, the set $\{h \in V : er(h) \leq er(h')\}$ is not removed from V. In particular, since we always have $er(h') \geq \nu$, by induction this implies the invariant $\inf_{h \in V} er(h) = \nu$, and therefore also

$$\forall t, er(h_t) - \nu = er(h_t) - \inf_{h \in V} er(h) = (er(h_t|R) - \inf_{h \in V} er(h|R))\mathbb{P}(R) \leq \beta_t,$$

where again the second equality is due to the fact that $\forall h \in V$, $\{x : h_t(x) \neq h(x)\} \subseteq DIS(V) \subseteq R$. We will spend the remainder of the proof bounding the size of n sufficient to guarantee some $\beta_t \leq \epsilon$. In particular, similar to the proof of Theorem 1, we will see that as long as $\beta_t > \epsilon$, we will halve $\mathbb{P}(DIS(V))$ roughly every $\tilde{O}\left(\theta^2 d\epsilon^{\frac{2}{\kappa} - 2}\right)$ label requests, so that the total number of label requests before some $\beta_t \leq \epsilon$ is at most roughly $\tilde{O}\left(\theta^2 d\epsilon^{\frac{2}{\kappa} - 2} \log(1/\epsilon)\right)$.

Recalling the definition of $h^{[k]}$ (from Definition 2), let

$$(4) \qquad V^{(\theta)} = \left\{ h \in V : \limsup_{k \to \infty} \mathbb{P}(h(X) \neq h^{[k]}(X)) > \frac{\mathbb{P}(R)}{2\theta} \right\}.$$

Note that after Step 7, if $V^{(\theta)} = \emptyset$, then

$\mathbb{P}(DIS(V))$

$$\leq \mathbb{P}\left(DIS\left(\left\{h \in \mathbb{C} : \limsup_{k \to \infty} \mathbb{P}\left(h(X) \neq h^{[k]}(X)\right) \leq \mathbb{P}(R)/(2\theta)\right\}\right)\right)$$

$$= \lim_{k' \to \infty} \mathbb{P}\left(DIS\left(\bigcap_{k > k'} B\left(h^{[k]}, \mathbb{P}(R)/(2\theta)\right)\right)\right)$$

$$\leq \lim_{k' \to \infty} \mathbb{P}\left(\bigcap_{k > k'} DIS\left(B\left(h^{[k]}, \mathbb{P}(R)/(2\theta)\right)\right)\right)$$

$$\leq \liminf_{k \to \infty} \mathbb{P}\left(DIS\left(B(h^{[k]}, \mathbb{P}(R)/(2\theta))\right)\right) \leq \liminf_{k \to \infty} \theta_{h^{[k]}} \frac{\mathbb{P}(R)}{2\theta} = \frac{\mathbb{P}(R)}{2},$$

so that we will satisfy the condition in Step 2 on the next round. Here we have used the definition of θ in the final inequality and equality. On the other hand, if after Step 7, we have $V^{(\theta)} \neq \emptyset$, then

$$\emptyset \neq \left\{h \in V : \limsup_{k \to \infty} \mathbb{P}(h(X) \neq h^{[k]}(X)) > \frac{\mathbb{P}(R)}{2\theta}\right\}$$

$$= \left\{h \in V : \left(\frac{\limsup_{k \to \infty} \mathbb{P}(h(X) \neq h^{[k]}(X))}{\mu}\right)^{\kappa} > \left(\frac{\mathbb{P}(R)}{2\mu\theta}\right)^{\kappa}\right\}$$

$$\subseteq \left\{h \in V : \left(\frac{diam(er(h) - \nu; \mathbb{C})}{\mu}\right)^{\kappa} > \left(\frac{\mathbb{P}(R)}{2\mu\theta}\right)^{\kappa}\right\}$$

$$\subseteq \left\{h \in V : er(h) - \nu > \left(\frac{\mathbb{P}(R)}{2\mu\theta}\right)^{\kappa}\right\}$$

$$= \left\{h \in V : er(h|R) - \inf_{h' \in V} er(h'|R) > \mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa}\right\}$$

$$\subseteq \left\{h \in V : UB(h, Q, \delta/n) - \min_{h' \in V} LB(h', Q, \delta/n) > \mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa}\right\}$$

$$\subseteq \left\{h \in V : LB(h, Q, \delta/n) - \min_{h' \in V} UB(h', Q, \delta/n)\right.$$
$$\left. > \mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa} - 4G(|Q|, \delta/n)\right\}.$$

Here, the third line follows from the fact that $er(h^{[k]}) \leq er(h)$ for all sufficiently large k, the fourth line follows from Condition 1, and the final line follows from the definition of UB and LB. By definition, every $h \in V$ has $LB(h, Q, \delta/n) \leq \min_{h' \in V} UB(h', Q, \delta/n)$, so for this last set to be nonempty after Step 7, we must have $\mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q|, \delta/n)$.

Combining these two cases ($V^{(\theta)} = \emptyset$ and $V^{(\theta)} \neq \emptyset$), since $|Q|$ gets reset to 0 upon reaching Step 3, we have that after every execution of Step 7,

$$\text{(5)} \qquad\qquad \mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q| - 1, \delta/n).$$

If $\mathbb{P}(R) \leq \frac{\epsilon}{2G(|Q|-1,\delta/n)} \leq \frac{\epsilon}{2G(|Q|,\delta/n)}$, then certainly $\beta_t \leq \epsilon$ (by definition of $\beta_t \leq 2G(|Q|, \delta/n)\mathbb{P}(R)$). So on any round for which $\beta_t > \epsilon$, we must have

$$\text{(6)} \qquad\qquad \frac{\epsilon}{2G(|Q| - 1, \delta/n)} < \mathbb{P}(R).$$

Combining (5) and (6), on any round for which $\beta_t > \epsilon$,

$$\text{(7)} \qquad \left(\frac{\epsilon}{2G(|Q| - 1, \delta/n)}\right)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q| - 1, \delta/n).$$

Solving for $G(|Q| - 1, \delta/n)$ reveals that when $\beta_t > \epsilon$,

$$\text{(8)} \qquad\qquad 4^{-1/\kappa}\left(\frac{\epsilon}{2}\right)^{\frac{\kappa-1}{\kappa}}(2\mu\theta)^{-1} < G(|Q| - 1, \delta/n).$$

Basic algebra shows that when $n \geq |Q| > d$, we have

$$G(|Q| - 1, \delta/n) \leq 3\sqrt{\frac{\ln\frac{4}{\delta} + (d+1)\ln(n)}{|Q|}}.$$

Combining this with (8), solving for $|Q|$, and adding d to handle the case $|Q| \leq d$, we have that on any round for which $\beta_t > \epsilon$,

$$\text{(9)} \qquad |Q| \leq \left(\frac{2}{\epsilon}\right)^{\frac{2\kappa-2}{\kappa}}(6\mu\theta)^2 4^{2/\kappa}\left(\ln\frac{4}{\delta} + (d+1)\ln(n)\right) + d.$$

Since $\beta_t \leq \mathbb{P}(R)$ by definition, and $\mathbb{P}(R)$ is at least halved each time we reach Step 3, we need to reach Step 3 at most $\lceil\log_2(1/\epsilon)\rceil$ times before we are guaranteed some $\beta_t \leq \epsilon$. Thus, any

$$\text{(10)} \qquad n \geq 1 + \left(\left(\frac{2}{\epsilon}\right)^{\frac{2\kappa-2}{\kappa}}(6\mu\theta)^2 4^{2/\kappa}\left(\ln\frac{4}{\delta} + (d+1)\ln(n)\right) + d\right)\log_2\frac{2}{\epsilon}$$

suffices to guarantee either some $|Q|$ exceeds (9) or we reach Step 3 at least $\lceil\log_2(1/\epsilon)\rceil$ times, either of which implies the existence of some $\beta_t \leq \epsilon$. The stated result now follows by basic inequalities to bound the smallest value of ϵ satisfying (10) for a given value of n. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

If the disagreement coefficient is finite, Theorem 4 can often represent a significant improvement in convergence rate compared to passive learning, where we typically expect rates of order $n^{-\kappa/(2\kappa-1)}$ [13, 27, 29]; this gap is especially notable when the disagreement coefficient and κ are small. Furthermore, the bound matches (up to logarithmic factors) the form of the minimax rate *lower bound* proved by Castro and Nowak [13] for threshold classifiers (where $\theta = 2$); as mentioned, that lower bound proof can be generalized to any nontrivial $\mathbb{C}$ (see Appendix D), so that the rate of Theorem 4 is nearly minimax optimal for any nontrivial $\mathbb{C}$ with *bounded* disagreement coefficients. Also note that, unlike the upper bound analysis of Castro and Nowak [13], we do not require the algorithm to be given any extra information about the noise distribution, so that this result is somewhat stronger; it is also more general, as this bound applies to an arbitrary hypothesis class.

A refined analysis and minor tweaks to the algorithm should be able to reduce the log factors in this result. For instance, defining UB and LB using the uniform convergence bounds of Alexander [1], and using a slightly more complicated algorithm closer to the original definition [6, 21] – taking multiple samples between bound evaluations, allowing a larger confidence argument to the UB and LB evaluations – the $\log^2 n$ factor should reduce at least to $\log n \log \log n$, if not further. Also, as previously mentioned, it is possible to replace the quantities $\mathbb{P}(R)$ and $\mathbb{P}(DIS(V))$ in Algorithm 1 with estimators of these quantities based on a finite sample of unlabeled data points, while preserving the results of Theorem 4 up to constant factors. For completeness, we provide an example of such estimators in Appendix C, along with a sketch of how the proof of Theorem 4 can be modified to compensate for using these estimated probabilities.

4.5. *Adaptive Rates in Active Learning: Algorithm 2.* Note that, as before, n gets divided by θ^2 in the rates achieved by Algorithm 1. As before, it is not clear whether any modification to the definitions of UB and LB can reduce this exponent on θ from 2 to 1. As such, it is natural to investigate the rates achieved by Algorithm 2 under Condition 1; we know that it does improve the dependence on θ for the worst case rates over distributions with any given noise rate, so we might hope that it does the same for the rates over distributions with any given values of μ and κ. Unfortunately, we do not presently know whether the original definition of Algorithm 2 achieves this improvement. However, we now present a slight modification of the algorithm, and prove that it does indeed provide the desired improvement in dependence on θ, while maintaining the improvements in the asymptotic dependence on n. Specifically, consider the following definition for the threshold in Algorithm 2.

$$\Delta_m(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta) = 3\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L} \cup Q, \delta; \mathcal{L}), \tag{11}$$

where $\hat{\mathcal{E}}_{\mathbb{C}}(\cdot, \cdot; \cdot)$ is defined in Appendix A, based on a notion of local Rademacher complexity studied by Koltchinskii [24]. In particular, the quantity $\hat{\mathcal{E}}_{\mathbb{C}}$ is known to be adaptive to Tsybakov's noise conditions, in the sense that more favorable noise conditions yield smaller values of $\hat{\mathcal{E}}_{\mathbb{C}}$. Using this definition, we have the following theorem; its proof is included in Appendix B.

THEOREM 5. *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 2 with threshold as in* (11)*, when allowed n label requests and given confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Condition 1 with finite parameter values κ and μ. Then there exists a finite (κ and μ -dependent) constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$
er(\hat{h}_n) - \nu \leq
\begin{cases}
c\left(d + \log \frac{1}{\delta}\right) \cdot exp\left\{-\sqrt{\frac{n}{c\theta(d + \log(1/\delta))}}\right\}, & when \ \kappa = 1 \\
c\left(\frac{\theta(d \log n + \log(1/\delta))}{n}\right)^{\frac{\kappa}{2\kappa-2}}, & when \ \kappa > 1
\end{cases}.
$$

◇

Note that this does indeed improve the dependence on θ, reducing its exponent from 2 to 1; we do lose some in that there is now a square root in the exponent of the $\kappa = 1$ case; however, as with Theorem 4, it is likely that slight refinements to the definition of Δ_m would reduce this (though we may also need to weaken the theorem statement to hold for any single n, rather than simultaneously for all n).

The bound in Theorem 5 is stated in terms of the VC dimension d. However, for certain nonparametric hypothesis classes, it is sometimes preferable to quantify the complexity of the class in terms of a constraint on the *entropy* of the class, relative to the distribution $\mathcal{D}_{XY}$ [see e.g., 13, 24, 29, 30]. Specifically, for $\epsilon \in [0, 1]$, define

$$
\omega_{\mathbb{C}}(m, \epsilon) = \mathbb{E} \sup_{\substack{h_1, h_2 \in \mathbb{C}: \\ \mathbb{P}\{h_1(X) \neq h_2(X)\} \leq \epsilon}} |(er(h_1) - er_m(h_1)) - (er(h_2) - er_m(h_2))|.
$$

CONDITION 2. *There exist finite constants $\alpha > 0$ and $\rho \in (0, 1)$ s.t. $\forall m \in \mathbb{N}$ and $\epsilon \in [0, 1]$, $\omega_{\mathbb{C}}(m, \epsilon) \leq \alpha \cdot \max\left\{\epsilon^{\frac{1-\rho}{2}} m^{-1/2}, m^{-\frac{1}{1+\rho}}\right\}$.* ◇

In particular, the entropy with bracketing condition used in the original minimax analysis of Tsybakov [29] implies Condition 2 [24], as does the analogous condition for random entropy [18, 19, 25]. In passive learning, it is known that empirical risk minimization achieves a rate of order $n^{-\kappa/(2\kappa+\rho-1)}$ under Conditions 1 *and* 2 [24, 25] (see also Appendix B, especially (19) and Lemma 5), and that this is sometimes minimax optimal [29]. The following theorem gives a bound on the rate of convergence of the same version of Algorithm 2 as in Theorem 5, this time in

terms of the entropy condition which, as before, is faster than the passive learning rate when the disagreement coefficient is finite. The proof of this result is included in Appendix B.

THEOREM 6. *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 2 with threshold as in* (11)*, when allowed n label requests and given confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Condition 1 with finite parameter values κ and μ, and Condition 2 with parameter values α and ρ. Then there exists a finite (κ, μ, α and ρ -dependent) constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu \leq c \left(\frac{\theta \log(n/\delta)}{n} \right)^{\frac{\kappa}{2\kappa + \rho - 2}} .$$

◇

Again, it is likely that refinements to the Δ_m definition may lead to improvements in the log factor. Also, although this result is stated for Algorithm 2, it is conceivable that, by modifying Algorithm 1 to use definitions of V and β_t based on $\hat{\mathcal{E}}_{\mathbb{C}}(Q, \delta; \emptyset)$, an analogous result might be possible for Algorithm 1 as well.

It is worth mentioning that Castro and Nowak [13] proved a minimax lower bound for the hypothesis class of *boundary fragments*, with an exponent having a similar dependence on related definitions of κ and ρ parameters to that of Theorem 6. Their result does provide a valid lower bound here; however, it is not clear whether their lower bound, Theorem 6, both, or neither is tight in the present context, since the value of θ is not presently known for that particular problem, and the matching upper bound of [13] was proven under a stronger restriction on the noise than Condition 1. However, see [35] for an analysis of the disagreement coefficient for other nonparametric hypothesis classes, characterized by smoothness of the decision surface.

5. Model Selection. While the previous sections address adaptation to the noise distribution, they are still restrictive in that they deal with hypothesis classes of limited expressiveness. That is, the assumption of finite VC dimension implies a strong restriction on the variety of classifiers one can represent (or approximate) in the class; the entropy conditions allow slightly more flexibility, but under nontrivial distributions, even the entropy conditions imply a significant restriction on the expressiveness of the class. Thus, for algorithms restricted to classifiers from such a restricted hypothesis class, it is often unrealistic to expect convergence to the Bayes error rate. We address this issue in this section by developing a general algorithm for learning with a sequence of nested hypothesis classes of increasing complexity, similar to the setting of Structural Risk Minimization in passive learning [31]. The objective is to adapt, not only to the noise conditions, but also to the complexity of

the optimal classifier. The starting point for this discussion is the assumption of a structure on $\mathbb{C}$, in the form of a sequence of nested hypothesis classes:

$$\mathbb{C}_1 \subset \mathbb{C}_2 \subset \cdots$$

Each class has an associated noise rate $\nu_i = \inf_{h \in \mathbb{C}_i} er(h)$, and we define $\nu_\infty = \lim_{i \to \infty} \nu_i$. We also let θ_i and d_i be the disagreement coefficient and VC dimension, respectively, for the set $\mathbb{C}_i$. We are interested in an algorithm that guarantees convergence in probability of the error rate to ν_∞. We are particularly interested in situations where $\nu_\infty = \nu^*$, a condition which is realistic in this setting since the sets $\mathbb{C}_i$ can be defined so that it is always satisfied, even while maintaining each $d_i < \infty$ [see e.g., 16]. Additionally, if we are so lucky as to have some $\nu_i = \nu^*$, then we would like the convergence rate achieved by the algorithm to be not significantly worse than running one of the above agnostic active learning algorithms with hypothesis class $\mathbb{C}_i$ alone. In this context, we can define a structure-dependent version of Tsybakov's noise condition as follows.

CONDITION 3. *For some nonempty $I \subseteq \mathbb{N}$, for each $i \in I$, there exist finite constants $\mu_i > 0$ and $\kappa_i \geq 1$, such that $\forall \epsilon > 0, diam(\epsilon; \mathbb{C}_i) \leq \mu_i \epsilon^{\frac{1}{\kappa_i}}$.* $\diamond$

Note that we do not require every $\mathbb{C}_i$, $i \in \mathbb{N}$, to have finite μ_i and κ_i, only some nonempty set $I \subseteq \mathbb{N}$; this is important, since we might not expect $\mathbb{C}_i$ to satisfy Condition 1 for small indices i, where the expressiveness is quite restricted.

In passive learning, there are several methods for this type of model selection which are known to preserve the convergence rates of each class $\mathbb{C}_i$ under Condition 3 [e.g., 24, 29]. In particular, Koltchinskii [24] develops a method that performs this type of model selection; it turns out we can modify Koltchinskii's method to suit our present needs in the context of active learning; this results in a general active learning model selection method that preserves the types of improved rates discussed in the previous section. This modification is presented below, based on using Algorithm 2 as a subroutine. (It should also be possible to define an analogous method that uses Algorithm 1 as a subroutine instead.)

Algorithm 3

Input: nested sequence of classes $\{\mathbb{C}_i\}$, label budget n, confidence parameter δ

Output: classifier $\hat{h}_n$

---

0. For $i = \lfloor \sqrt{n/2} \rfloor, \lfloor \sqrt{n/2} \rfloor - 1, \lfloor \sqrt{n/2} \rfloor - 2, \ldots, 1$
1.   Let $\mathcal{L}_{in}$ and Q_{in} be the sets returned by Algorithm 2 run with $\mathbb{C}_i$ and the threshold (11), allowing $\lfloor n/(2i^2) \rfloor$ label requests, and confidence $\delta/(2i^2)$
2.   Let $h_{in} \leftarrow \text{LEARN}_{\mathbb{C}_i}(\cup_{j \geq i} \mathcal{L}_{jn}, Q_{in})$
3.   If $h_{in} \neq \varnothing$ and $\forall j$ s.t. $i < j \leq \lfloor \sqrt{n/2} \rfloor$,
    $$er_{\mathcal{L}_{jn} \cup Q_{jn}}(h_{in}) - er_{\mathcal{L}_{jn} \cup Q_{jn}}(h_{jn}) \leq \tfrac{3}{2} \hat{\mathcal{E}}_{\mathbb{C}_j}(\mathcal{L}_{jn} \cup Q_{jn}, \delta/(2j^2); \mathcal{L}_{jn})$$
4. $\hat{h}_n \leftarrow h_{in}$
5. Return $\hat{h}_n$

The function $\hat{\mathcal{E}}.(\cdot, \cdot; \cdot)$ is defined in Appendix A. This method can be shown to have a confidence bound on its error rate converging to ν_∞ at a rate never significantly worse than the original passive learning method of Koltchinskii [24], as desired. Additionally, we have the following guarantee on the rate of convergence under Condition 3. The proof is similar in style to Koltchinskii's original proof, though some care is needed due to the altered sampling distribution and the constraint set $\mathcal{L}_{jn}$. The proof is included in Appendix B.

THEOREM 7. *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 3, when allowed n label requests and confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Condition 3. Then there exist finite (κ_i and μ_i -dependent) constants c_i such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu_\infty \leq 3\min_{i \in I}(\nu_i - \nu_\infty) + \begin{cases} c_i\left(d_i + \log \tfrac{1}{\delta}\right) \cdot exp\left\{-\sqrt{\frac{n}{c_i \theta_i(d_i + \log \frac{1}{\delta})}}\right\}, & \text{if } \kappa_i = 1 \\ c_i\left(\frac{\theta_i(d_i \log n + \log \frac{1}{\delta})}{n}\right)^{\frac{\kappa_i}{2\kappa_i - 2}}, & \text{if } \kappa_i > 1 \end{cases}.$$

$\diamond$

In particular, if we are so lucky as to have $\nu_i = \nu^*$ for some finite i, then the above algorithm achieves a convergence rate not significantly worse than that guaranteed by Theorem 5 for applying Algorithm 2 directly, with hypothesis class $\mathbb{C}_i$. Note that the algorithm itself has no dependence on the set I, nor has it any dependence on each class's complexity parameters $d_i, \kappa_i, \mu_i, \theta_i$; the adaptive behavior of the data-dependent bound $\hat{\mathcal{E}}_{\mathbb{C}_j}$ allows the algorithm to adaptively ignore the returned classifier from the runs of Algorithm 2 for which convergence is slow, thus automatically selecting an index for which the error rate is relatively small.

As in the previous section, we can also show a variant of this result when the complexities are quantified in terms of the entropy. Specifically, consider the following condition and theorem; the proof is in Appendix B. Again, this represents an improvement over known results for passive learning when the disagreement coefficients are finite.

CONDITION 4.   *For each $i \in \mathbb{N}$, there exist finite constants $\alpha_i > 0$, $\rho_i \in (0, 1)$ s.t. $\forall m \in \mathbb{N}$ and $\epsilon \in [0, 1]$, $\omega_{\mathbb{C}_i}(m, \epsilon) \leq \alpha_i \cdot \max \left\{ \epsilon^{\frac{1-\rho_i}{2}} m^{-1/2}, m^{-\frac{1}{1+\rho_i}} \right\}.$*   ◇

THEOREM 8.   *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 3, when allowed n label requests and confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Conditions 3 and 4. Then there exist finite (κ_i, μ_i, α_i and ρ_i -dependent) constants c_i such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu_\infty \leq 3 \min_{i \in I}(\nu_i - \nu_\infty) + c_i \left(\frac{\theta_i \log(n/\delta)}{n} \right)^{\frac{\kappa_i}{2\kappa_i + \rho_i - 2}}.$$

◇

In addition to these theorems for this structure-dependent version of Tsybakov's noise conditions, we also have the following result for a structure-independent noise condition, in the sense that the noise condition does not depend on the particular choice of $\mathbb{C}_i$ sets, but only on the distribution $\mathcal{D}_{XY}$ (and in some sense, the full class $\mathbb{C} = \bigcup_i \mathbb{C}_i$); it may be particularly useful when the class $\mathbb{C}$ is universal, in the sense that it can approximate any classifier.

THEOREM 9.   *Suppose the sequence $\{\mathbb{C}_i\}$ is constructed so that $\nu_\infty = \nu^*$, and $\hat{h}_n$ is the classifier returned by Algorithm 3, when allowed n label requests and confidence parameter $\delta \in (0, 1/2)$. Suppose that there exists a constant $\mu > 0$ s.t. for all measurable $h : \mathcal{X} \rightarrow \mathcal{Y}$, $er(h) - \nu^* \geq \mu\mathbb{P}\{h(X) \neq h^*(X)\}$. Then there exists a finite (μ-dependent) constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu^* \leq c \min_{i \in \mathbb{N}}(\nu_i - \nu^*) + \left(d_i + \log \frac{i}{\delta} \right) \cdot exp \left\{ -\sqrt{\frac{n}{ci^2\theta_i(d_i + \log(i/\delta))}} \right\}.$$

◇

The condition $\nu_\infty = \nu^*$ is quite easy to satisfy: for examples, $\mathbb{C}_i$ could be axis-aligned decision trees of depth i, or thresholded polynomials of degree i, or multi-layer neural networks with i internal units, etc. As for the noise condition in Theorem 9, this would be satisfied whenever $\mathbb{P}(|\eta(X) - 1/2| \geq c) = 1$ for some constant $c \in (0, 1/2]$. The case where $er(h) - \nu^* \geq \mu\mathbb{P}\{h(X) \neq h^*(X)\}^\kappa$ for $\kappa > 1$ can be studied analogously, though the rate improvements over passive learning are more subtle.

6. Conclusions. Under Tsybakov's noise conditions, active learning can offer improved asymptotic convergence rates compared to passive learning when the disagreement coefficient is finite. It is also possible to preserve these improved convergence rates when learning with a nested structure of hypothesis classes, using an algorithm that adapts to both the noise conditions and the complexity of the optimal classifier.

APPENDIX A: DEFINITION OF $\hat{\mathcal{E}}$ AND RELATED QUANTITIES

We define the quantity $\hat{\mathcal{E}}_{\mathbb{C}}$ following Koltchinskii's analysis of excess risk in terms of local Rademacher complexity [24]. The general idea is to construct a bound on the excess risk achieved by a given algorithm, such as empirical risk minimization, via an application of Talagrand's inequality. Such a bound should be based on a measure of the expressiveness of the set of functions $\mathbb{C}$; however, to bound the excess risk achieved by a particular algorithm given a number of data points, we need only measure the expressiveness of the set of functions the algorithm is likely to select from. For reasonable algorithms, such as empirical risk minimization, this means the set of functions with reasonably small excess risk. Thus, we can bound the excess risk of the algorithm in terms of a measure of expressiveness of the set of functions with relatively small risk, typically referred to as a *local* complexity measure. This reasoning is somewhat circular, in that first we must decide how small to expect the excess risk of the returned function to be before we can calculate the local complexity measure, which itself is used to calculate a bound on the risk of the returned function. Thus, we define the bound on the excess risk as a kind of fixed point. Furthermore, we can estimate these quantities using data-dependent confidence bounds, so that the excess risk bound can be calculated without direct access to the distribution. For the data-dependent measure of the expressiveness of the function class, we can use a Rademacher process. A detailed motivation and derivation can be found in [24].

For our purposes, we add an additional constraint, by requiring the functions we calculate the complexity of to agree with the labels of a labeled set $\mathcal{L}$. This is helpful for us, since given a set Q of labeled data with true labels, for any two functions h_1 and h_2 that agree on the labels of $\mathcal{L}$, it is always true that $er_{\mathcal{L} \cup Q}(h_1) - er_{\mathcal{L} \cup Q}(h_2)$ equals the difference of the true empirical error rates. As we prove below, as long as the set $\mathcal{L}$ is chosen carefully (i.e., as in Algorithm 2), the addition of this constraint is essentially inconsequential, so that $\hat{\mathcal{E}}_{\mathbb{C}}$ remains a valid excess risk bound. The detailed definitions are stated as follows.

For any function $f : \mathcal{X} \to \mathbb{R}$, and $\xi_1, \xi_2, \ldots$ a sequence of independent random variables with distribution uniform in $\{-1, +1\}$, define the *Rademacher process*

for f under a finite set of (index, label) pairs $S \subset \mathbb{N} \times \mathcal{Y}$ as

$$R(f; S) = \frac{1}{|S|} \sum_{(i,y) \in S} \xi_i f(X_i).$$

The ξ_i should be thought of as internal variables in the learning algorithm, rather than being fundamental to the learning problem.

For any two finite sets $\mathcal{L} \subset \mathbb{N} \times \mathcal{Y}$ and $S \subset \mathbb{N} \times \mathcal{Y}$, define

$$\mathbb{C}[\mathcal{L}] = \{h \in \mathbb{C} : er_\mathcal{L}(h) = 0\},$$

$$\hat{\mathbb{C}}(\epsilon; \mathcal{L}, S) = \{h \in \mathbb{C}[\mathcal{L}] : er_S(h) - \min_{h' \in \mathbb{C}[\mathcal{L}]} er_S(h') \leq \epsilon\},$$

$$\hat{D}_\mathbb{C}(\epsilon; \mathcal{L}, S) = \sup_{h_1, h_2 \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}, S)} \frac{1}{|S|} \sum_{(i,y) \in S} \mathbb{1}[h_1(X_i) \neq h_2(X_i)],$$

and $\quad \hat{\phi}_\mathbb{C}(\epsilon; \mathcal{L}, S) = \frac{1}{2} \sup_{h_1, h_2 \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}, S)} R(h_1 - h_2; S).$

For $\delta, \epsilon > 0$, $m \in \mathbb{N}$, define $s_m(\delta) = \ln \frac{20m^2 \log_2(3m)}{\delta}$ and $\mathbb{Z}_\epsilon = \{j \in \mathbb{Z} : 2^j \geq \epsilon\}$, and for any set $S \subset \mathbb{N} \times \mathcal{Y}$, define the set $S^{(m)} = \{(i, y) \in S : i \leq m\}$. We use the following definitions from Koltchinskii [24] with only minor modifications.

DEFINITION 3. *For $\epsilon \in [0, 1]$, and finite sets $S, \mathcal{L} \subset \mathbb{N} \times \mathcal{Y}$, define*

$$\hat{U}_\mathbb{C}(\epsilon, \delta; \mathcal{L}, S) = \hat{K} \left(\hat{\phi}_\mathbb{C}(\hat{c}\epsilon; \mathcal{L}, S) + \sqrt{\frac{s_{|S|}(\delta) \hat{D}_\mathbb{C}(\hat{c}\epsilon; \mathcal{L}, S)}{|S|}} + \frac{s_{|S|}(\delta)}{|S|} \right),$$

and $\quad \hat{\mathcal{E}}_\mathbb{C}(S, \delta; \mathcal{L}) = \inf \left\{ \epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \min_{m \in \mathbb{N}} \hat{U}_\mathbb{C}(2^j, \delta; \mathcal{L}^{(m)}, S^{(m)}) \leq 2^{j-4} \right\},$

where, for our purposes, we can take $\hat{K} = 752$, and $\hat{c} = 3/2$, though there seems to be room for improvement in these constants. For completeness, we also define $\hat{\mathcal{E}}_\mathbb{C}(\emptyset, \delta; \mathbb{C}, \mathcal{L}) = \infty$ by convention. $\diamond$

We also define a related quantity, representing a distribution-dependent version of $\hat{\mathcal{E}}$, also explored by Koltchinskii [24]. For $\epsilon > 0$, define

$$\mathbb{C}(\epsilon) = \{h \in \mathbb{C} : er(h) - \nu \leq \epsilon\}.$$

For $m \in \mathbb{N}$, let

$$\phi_{\mathbb{C}}(m, \epsilon) = \mathbb{E} \sup_{h_1, h_2 \in \mathbb{C}(\epsilon)} |(er(h_1) - er_m(h_1)) - (er(h_2) - er_m(h_2))|,$$

$$\tilde{U}_{\mathbb{C}}(m, \epsilon, \delta) = \tilde{K} \left(\phi_{\mathbb{C}}(m, \tilde{c}\epsilon) + \sqrt{\frac{s_m(\delta) diam(\tilde{c}\epsilon; \mathbb{C})}{m}} + \frac{s_m(\delta)}{m} \right),$$

and $\quad \tilde{\mathcal{E}}_{\mathbb{C}}(m, \delta) = \inf \left\{ \epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \tilde{U}_{\mathbb{C}}(m, 2^j, \delta) \leq 2^{j-4} \right\},$

where, for our purposes, we can take $\tilde{K} = 8272$ and $\tilde{c} = 3$. For completeness, we also define $\tilde{\mathcal{E}}_{\mathbb{C}}(0, \delta) = \infty$.

A.1. Definition of r_0. In Definition 1, we took $r_0 = 0$. If $\theta < \infty$, then this choice is usually relatively harmless. However, in some cases, setting $r_0 = 0$ results in a suboptimal, or even infinite, value of θ, which is undesirable. In these cases, we would like to set r_0 as large as possible while maintaining the validity of the bounds. If we do this carefully enough, we should be able to establish bounds that, even in the worst case when $\theta = 1/r_0$, are never worse than the bounds for some analogous passive learning method; however, to do this requires r_0 to depend on the parameters of the learning problem: namely, n, δ, $\mathbb{C}$, and $\mathcal{D}_{XY}$. The effect of a larger r_0 can sometimes be dramatic, as there are scenarios where $1 \ll \theta \ll 1/r_0$ [8]; we certainly wish to distinguish between such scenarios, and those where $\theta \propto 1/r_0$.

Generally, depending on the bound we wish to prove, different values of r_0 may be appropriate. For the tightest bound in terms of θ proven below (namely, Lemma 7), the definition of $r_0 = r_{\mathbb{C}}(n, \delta)$ in (13) below gives a good bound. For the looser bounds (namely, Theorems 5 and 6), a larger value of r_0 may provide better bounds; however, this same general technique can be employed to define a good value for r_0 in these looser bounds as well, simply using upper bounds on (13) analogous to how the theorems themselves are derived from Lemma 7 in their proof below. Likewise, one can state analogous refinements of r_0 for Theorems 1 - 4, though for brevity these are left for the reader's independent consideration.

DEFINITION 4. *Define*
(12)
$$\tilde{m}_{\mathbb{C}}(n, \delta) = \min \left\{ m \in \mathbb{N} : n \leq \log_2 \frac{4m^2}{\delta} + 2e \sum_{\ell=0}^{m-1} \mathbb{P} \left(DIS \left(\mathbb{C}(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)) \right) \right) \right\}$$

and

(13) $$r_{\mathbb{C}}(n, \delta) = \max \left\{ \frac{1}{\tilde{m}_{\mathbb{C}}(n, \delta)} \sum_{\ell=0}^{\tilde{m}_{\mathbb{C}}(n,\delta)-1} diam(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \mathbb{C}), 2^{-n} \right\}.$$

$\diamond$

We use this definition of $r_0 = r_{\mathbb{C}}(n, \delta)$ in all of the proofs below. In partic-ular, with this definition, Lemma 7 is never significantly worse than the analo-gous known result for passive learning (though it can be significantly better when $\theta << 1/r_0$).

APPENDIX B: MAIN PROOFS

Recall that $\mathcal{Z}^{(m)} = \{(i, Y_i) : i \leq m\}$ is the indexed *true* labels for the first m points in the data sequence. Let $\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta) = \hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{Z}^{(m)}, \delta; \emptyset)$. For each $m \in \mathbb{N}$, let $\hat{h}_m^* = \underset{h \in \mathbb{C}}{\operatorname{argmin}}\, er_m(h)$ be the empirical risk minimizer in $\mathbb{C}$ for the *true* labels of the first m data points. The following lemma is crucial to all of the proofs that follow.

LEMMA 5. *[24] For $\delta \in (0, 1/2)$, there is an event $E_{\mathbb{C}, \delta}$ with $\mathbb{P}(E_{\mathbb{C}, \delta}) \geq 1 - \delta/2$ such that, on event $E_{\mathbb{C}, \delta}$, $\forall m \in \mathbb{N}, \forall h \in \mathbb{C}, \forall \tau \in (0, 1/m), \forall h' \in \mathbb{C}(\tau)$,*

$$er(h) - \nu \leq \max\left\{2(er_m(h) - er_m(h') + \tau), \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)\right\}$$

$$er_m(h) - er_m(\hat{h}_m^*) \leq \tfrac{3}{2}\max\left\{(er(h) - \nu), \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)\right\},$$

$$\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta) \leq \tilde{\mathcal{E}}_{\mathbb{C}}(m, \delta),$$

and for any $j \in \mathbb{Z}$ with $2^j > \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$,

$$\sup_{h_1, h_2 \in \mathbb{C}(2^j)} |(er_m(h_1) - er(h_1)) - (er_m(h_2) - er(h_2))| \leq \hat{U}_{\mathbb{C}}(2^j, \delta; \emptyset, \mathcal{Z}^{(m)}).$$

$\diamond$

This lemma essentially follows from details of the proof of Koltchinskii's The-orem 1, Lemma 2, and Theorem 3 [24], combined with a union bound so that the results hold simultaneously for all m. We will not prove Lemma 5 here. The reader is referred to Koltchinskii's paper for the details. Specifically, each of the four inequalities of Lemma 5 basically follows from Step 5 of Koltchinskii's proof of his Theorem 3 in [24], in combination with other facts. In particular, the third inequality in Lemma 5 directly follows from Koltchinskii's Theorem 3. The fourth inequality follows from a combination of the aforementioned Step 5, and bounds on the probability of a particular event $E_n(s_n(\delta))$ given in the proof of Koltchin-skii's Theorem 1. The second inequality in Lemma 5 follows from a combination of Koltchinskii's Theorem 3 and (9.2) in his Lemma 2 proof. Note that the addition of the "$\underset{m \in \mathbb{N}}{\min}$" in our definition of $\hat{\mathcal{E}}_{\mathbb{C}}$, compared to Koltchinskii's original definition

(of $\hat{\delta}_n(t)$), does not cause any problems for Theorem 3 in [24], since we are employing a union bound to enable Koltchinskii's Theorem 3 to apply simultaneously for all n anyway, and since $\phi_{\mathbb{C}}(n, \epsilon)$ and $\frac{s_n(\delta)}{n}$ are nonincreasing functions of n. The first inequality in our Lemma 5 follows from Koltchinskii's Theorem 3, combined with a slight twist on the sequence of bounds on the excess risk found in the middle of Koltchinskii's page 2633; specifically, we can obtain this first inequality if we follow this same sequence of bound relaxations, except *not* relaxing the difference of empirical risks to the excess empirical risk (part of the third relaxation in the sequence). Finally, we have set $s_n(\delta)$ precisely so that Koltchinskii's lower bounds on the probabilities of the relevant events imply that all four inequalities in our Lemma 5 hold for all m, h, τ, and h' with the desired $1 - \delta/2$ probability.

B.1. Proofs Relating to Section 4. For the proofs that follow, we will let $\mathcal{L}$ and Q be the sets returned by Algorithm 2. It is important to keep in mind that, as defined, the sets $\mathcal{L}^{(m)}$ and $Q^{(m)}$ are indeed the values that the sets $\mathcal{L}$ and Q have at the conclusion of round m in the algorithm execution (i.e., after processing the data point X_m). Furthermore, every index processed is added to either $\mathcal{L}$ or Q, so that we always have $|\mathcal{L}^{(m)}| + |Q^{(m)}| = |\mathcal{L}^{(m)} \cup Q^{(m)}| = m$, for any $m \leq |\mathcal{L} \cup Q|$.

LEMMA 6. *On event $E_{\mathbb{C},\delta}$ (of Lemma 5), $\forall m \in \mathbb{N} \cup \{0\}$ with $m \leq |\mathcal{L} \cup Q|$,*

$$\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}^{(m)} \cup Q^{(m)}, \delta; \mathcal{L}^{(m)}) = \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta),$$

$$\forall \epsilon \geq 0, \ \hat{h}_m^* \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) \subseteq \hat{\mathbb{C}}(\epsilon; \emptyset, \mathcal{Z}^{(m)}),$$

and

$$\forall \epsilon \leq \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta), \ \hat{U}_{\mathbb{C}}(\epsilon, \delta; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) = \hat{U}_{\mathbb{C}}(\epsilon, \delta; \emptyset, \mathcal{Z}^{(m)}).$$

◇

PROOF OF LEMMA 6. Throughout this proof, we assume the event $E_{\mathbb{C},\delta}$ occurs. We proceed by induction on m, with the base case of $m = 0$ (which clearly holds). Suppose the statements are true for all $m' < m$. The case $\mathcal{L}^{(m)} = \emptyset$ is trivial, so assume $\mathcal{L}^{(m)} \neq \emptyset$. For the inductive step, suppose $\epsilon \leq \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$, and take any

$$h \in \hat{\mathbb{C}}(\epsilon; \emptyset, \mathcal{Z}^{(m)}).$$

In particular, by Lemma 5, taking any $\tau \in (0, 1/m)$ and $h' \in \mathbb{C}(\tau)$, we have

$$er(h) - \nu \leq \max \left\{ 2(er_m(h) - er_m(h') + \tau), \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta) \right\}$$

$$\leq \max \left\{ 2(er_m(h) - er_m(\hat{h}_m^*) + \tau), \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta) \right\}.$$

Taking the limit as $\tau \to 0$ implies

$$er(h) - \nu \leq \max\left\{2(er_m(h) - er_m(\hat{h}_m^*)), \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)\right\} \leq 2\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta),$$

and thus for any non-negative integer $m' < m$ and any $h' \in \mathbb{C}$, Lemma 5 implies

$$er_{m'}(h) - er_{m'}(h') \leq er_{m'}(h) - er_{m'}(\hat{h}_{m'}^*) \leq \frac{3}{2}\max\left\{er(h) - \nu, \hat{\mathcal{E}}_{\mathbb{C}}(m', \delta)\right\}$$

$$\leq \frac{3}{2}\max\left\{2\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta), \hat{\mathcal{E}}_{\mathbb{C}}(m', \delta)\right\} \leq 3\hat{\mathcal{E}}_{\mathbb{C}}(m', \delta) = 3\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}^{(m')} \cup Q^{(m')}, \delta; \mathcal{L}^{(m')}).$$

Since this is below the threshold $\Delta_{m'}(\mathcal{L}^{(m')}, Q^{(m')}, h, h', \delta)$, by induction on $m' < m$ we must have $er_{\mathcal{L}^{(m)}}(h) = 0$, and therefore $h \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)})$. Since this is the case for all such h, we must have that

(14) $$\hat{\mathbb{C}}(\epsilon; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) \supseteq \hat{\mathbb{C}}(\epsilon; \emptyset, \mathcal{Z}^{(m)}).$$

In particular, this implies that

$$\hat{U}_{\mathbb{C}}(\epsilon, \delta; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) \geq \hat{U}_{\mathbb{C}}(\epsilon, \delta; \emptyset, \mathcal{Z}^{(m)}).$$

Since $\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$ is non-increasing in m, by the inductive hypothesis, this means

$$\min_{m' \leq m} \hat{U}_{\mathbb{C}}(\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta), \delta; \mathcal{L}^{(m')}, \mathcal{L}^{(m')} \cup Q^{(m')})$$

$$\geq \min_{m' \leq m} \hat{U}_{\mathbb{C}}(\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta), \delta; \emptyset, \mathcal{Z}^{(m')}) > \frac{1}{16}\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta),$$

where the last inequality follows from the definition of $\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$, (which is a power of 2). Thus, we must have $\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}^{(m)} \cup Q^{(m)}, \delta; \mathcal{L}^{(m)}) \geq \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$.

The relation in (14) also implies that

$$\hat{h}_m^* \in \hat{\mathbb{C}}(\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta); \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) \subseteq \mathbb{C}[\mathcal{L}^{(m)}],$$

and therefore

$$\forall \epsilon \geq 0, \ \hat{h}_m^* \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}^{(m)}, \mathcal{L}^{(m)} \cup Q^{(m)}) \subseteq \hat{\mathbb{C}}(\epsilon; \emptyset, \mathcal{Z}^{(m)}).$$

The inductive hypothesis already gives us that

$$\forall \epsilon \geq 0, \forall m' < m, \hat{\mathbb{C}}(\epsilon; \mathcal{L}^{(m')}, \mathcal{L}^{(m')} \cup Q^{(m')}) \subseteq \hat{\mathbb{C}}(\epsilon; \emptyset, \mathcal{Z}^{(m')}).$$

Combining these facts implies

$$\forall \epsilon \geq 0, \forall m' \leq m, \ \hat{U}_{\mathbb{C}}(\epsilon, \delta; \mathcal{L}^{(m')}, \mathcal{L}^{(m')} \cup Q^{(m')}) \leq \hat{U}_{\mathbb{C}}(\epsilon, \delta; \emptyset, \mathcal{Z}^{(m')}),$$

and therefore

$$\forall \epsilon \geq 0, \min_{m' \leq m} \hat{U}_{\mathbb{C}}(\epsilon, \delta; \mathcal{L}^{(m')}, \mathcal{L}^{(m')} \cup Q^{(m')}) \leq \min_{m' \leq m} \hat{U}_{\mathbb{C}}(\epsilon, \delta; \emptyset, \mathcal{Z}^{(m')}).$$

But this means $\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}^{(m)} \cup Q^{(m)}, \delta; \mathcal{L}^{(m)}) \leq \hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$. The lemma now follows by the principle of induction. $\qquad\square$

LEMMA 7.   *Suppose for any $n \in \mathbb{N}$, $\hat{h}_n$ is the classifier returned by Algorithm 2 with threshold as in* (11)*, when allowed n label requests and given confidence parameter $\delta \in (0, 1/2)$, and suppose further that m_n is the value of $|Q| + |\mathcal{L}|$ when Algorithm 2 returns. Then there is an event $H_{\mathbb{C},\delta}$ such that $\mathbb{P}(H_{\mathbb{C},\delta} \cap E_{\mathbb{C},\delta}) \geq 1 - \delta$, such that on $H_{\mathbb{C},\delta} \cap E_{\mathbb{C},\delta}$, $\forall n \in \mathbb{N}$,*

$$(15) \qquad\qquad er(\hat{h}_n) - \nu \leq \tilde{\mathcal{E}}_{\mathbb{C}}(m_n, \delta),$$

and

$$(16) \qquad n \leq \min\left\{ m_n, \log_2 \frac{4m_n^2}{\delta} + 4e\theta \left(1 + \sum_{\ell=0}^{m_n-1} diam(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \mathbb{C}) \right) \right\}. \qquad\diamond$$

PROOF OF LEMMA 7.   Again, assume event $E_{\mathbb{C},\delta}$ occurs. By Lemma 5, $\forall \tau \in (0, 1/m_n)$ and $h'_n \in \mathbb{C}(\tau)$,

$$er(\hat{h}_n) - \nu \leq \max\left\{ 2(er_{m_n}(\hat{h}_n) - er_{m_n}(h'_n) + \tau), \hat{\mathcal{E}}_{\mathbb{C}}(m_n, \delta) \right\}$$
$$\leq \max\left\{ 2(er_{m_n}(\hat{h}_n) - er_{m_n}(\hat{h}^*_{m_n}) + \tau), \hat{\mathcal{E}}_{\mathbb{C}}(m_n, \delta) \right\}.$$

Letting $\tau \to 0$, and noting that $er_{\mathcal{L}}(\hat{h}^*_{m_n}) = 0$ (Lemma 6) implies $er_{m_n}(\hat{h}_n) = er_{m_n}(\hat{h}^*_{m_n})$, we have

$$er(\hat{h}_n) - \nu \leq \hat{\mathcal{E}}_{\mathbb{C}}(m_n, \delta) \leq \tilde{\mathcal{E}}_{\mathbb{C}}(m_n, \delta),$$

where the last inequality is also due to Lemma 5. Note that this $\hat{\mathcal{E}}_{\mathbb{C}}(m_n, \delta)$ represents an interesting data-dependent bound.

To get the bound on the number of label requests, we proceed as follows. For any $m \in \mathbb{N}$, and integer $\ell \in [0, m)$, let I_ℓ be the indicator for the event that Algorithm 2 requests the label $Y_{\ell+1}$ and let $N_m = \sum_{\ell=0}^{m-1} I_\ell$. Also, let I'_ℓ be the indicator for the event $\{X_{\ell+1} \in DIS(\mathbb{C}(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)))\}$, and let $N'_m = \sum_{\ell=0}^{m-1} I'_\ell$. Noting that

$$I_\ell = \mathbb{1}\left[\{X_{\ell+1} \in DIS(\hat{\mathbb{C}}(3\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}^{(\ell)} \cup Q^{(\ell)}, \delta; \mathcal{L}^{(\ell)}); \mathcal{L}^{(\ell)}, \mathcal{L}^{(\ell)} \cup Q^{(\ell)}))\} \cap \{m_n > \ell\} \right],$$

we have that for any $q \geq 0$,

$$
\begin{aligned}
& \mathbb{P}\left[\{N_m > q\} \cap E_{\mathbb{C},\delta}\right] \\
& \leq \mathbb{P}\left[\sum_{\ell=0}^{m-1} \mathbb{1}\left[\{X_{\ell+1} \in DIS(\hat{\mathbb{C}}(3\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta); \emptyset, \mathcal{Z}^{(\ell)}))\} \cap E_{\mathbb{C},\delta}\right] > q\right] \\
& \leq \mathbb{P}\left[\sum_{\ell=0}^{m-1} \mathbb{1}\left[\{X_{\ell+1} \in DIS(\mathbb{C}(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta)))\}\right] > q\right] = \mathbb{P}\left[N_m' > q\right].
\end{aligned}
$$

The first inequality is due to Lemmas 6 and 5, while the second inequality is due to Lemma 5. Note that

$$
\mathbb{E}[N_m'] = \sum_{\ell=0}^{m-1} \mathbb{P}[I_\ell' = 1] = \sum_{\ell=0}^{m-1} \mathbb{P}\left(DIS\left(\mathbb{C}(2\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta))\right)\right).
$$

Let us name this last quantity q_m. Thus, by union and Chernoff bounds,

$$
\begin{aligned}
& \mathbb{P}\left[\left\{\exists m \in \mathbb{N} : N_m > \max\left\{2eq_m, q_m + \log_2 \frac{4m^2}{\delta}\right\}\right\} \cap E_{\mathbb{C},\delta}\right] \\
& \leq \sum_{m \in \mathbb{N}} \mathbb{P}\left[\left\{N_m > \max\left\{2eq_m, q_m + \log_2 \frac{4m^2}{\delta}\right\}\right\} \cap E_{\mathbb{C},\delta}\right] \\
& \leq \sum_{m \in \mathbb{N}} \mathbb{P}\left[\left\{N_m' > \max\left\{2eq_m, q_m + \log_2 \frac{4m^2}{\delta}\right\}\right\}\right] \leq \sum_{m \in \mathbb{N}} \frac{\delta}{4m^2} \leq \frac{\delta}{2}.
\end{aligned}
$$

For any n, we know $n \leq m_n \leq 2^n$. Therefore, we have that on an event (which includes $E_{\mathbb{C},\delta}$) occurring with probability $\geq 1 - \delta$, for every $n \in \mathbb{N}$,

$$
\begin{aligned}
n \leq \max\{N_{m_n}, \log_2 m_n\} &\leq \max\left\{2eq_{m_n}, q_{m_n} + \log_2 \frac{4m_n^2}{\delta}\right\} \\
&\leq \log_2 \frac{4m_n^2}{\delta} + 2e \sum_{\ell=0}^{m_n-1} \mathbb{P}\left(DIS\left(\mathbb{C}(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta))\right)\right).
\end{aligned}
$$

In particular, this implies $\tilde{m}_n := \tilde{m}_{\mathbb{C}}(n,\delta) \leq m_n$ (where $\tilde{m}_{\mathbb{C}}(n,\delta)$ is defined in (12)). We now use the definition of θ with the r_0 in (13).

$$n \leq \log_2 \frac{4\tilde{m}_n^2}{\delta} + 2e \sum_{\ell=0}^{\tilde{m}_n-1} \mathbb{P}\bigg(DIS\Big(\mathbb{C}(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta))\Big)\bigg)$$

$$\leq \log_2 \frac{4\tilde{m}_n^2}{\delta} + 2e\theta \sum_{\ell=0}^{\tilde{m}_n-1} \max \Big\{ diam(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta); \mathbb{C}), r_{\mathbb{C}}(n,\delta) \Big\}$$

$$\leq \log_2 \frac{4\tilde{m}_n^2}{\delta} + 4e\theta \left(1 + \sum_{\ell=0}^{\tilde{m}_n-1} diam(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta); \mathbb{C})\right)$$

$$\leq \log_2 \frac{4m_n^2}{\delta} + 4e\theta \left(1 + \sum_{\ell=0}^{m_n-1} diam(6\tilde{\mathcal{E}}_{\mathbb{C}}(\ell,\delta); \mathbb{C})\right).$$

<div align="right">□</div>

LEMMA 8.   *On event $H_{\mathbb{C},\delta} \cap E_{\mathbb{C},\delta}$, under Condition 1, $\forall n \in \mathbb{N}$,*

$$\min \Big\{ \tilde{\mathcal{E}}_{\mathbb{C}}(m_n, \delta), 1 \Big\} \leq \begin{cases} c \left(d + \log \frac{1}{\delta}\right) \cdot exp\left\{ -\sqrt{\frac{n}{c\theta(d+\log(1/\delta))}}\right\}, & if\ \kappa = 1 \\ c \left(\frac{\theta(d\log n+\log(1/\delta))}{n}\right)^{\frac{\kappa}{2\kappa-2}}, & if\ \kappa > 1 \end{cases},$$

for a finite constant c (depending on κ and μ); under the additional Condition 2, $\forall n \in \mathbb{N}$,

$$\min \Big\{ \tilde{\mathcal{E}}_{\mathbb{C}}(m_n, \delta), 1 \Big\} \leq c \left(\frac{\theta \log(n/\delta)}{n}\right)^{\frac{\kappa}{2\kappa+\rho-2}},$$

for a finite constant c (depending on κ, μ, ρ, and α).   ◇

PROOF OF LEMMA 8.   We begin with the first case (Condition 1 only). We know that for $\epsilon \in (0, 1)$,

$$\omega_{\mathbb{C}}(m, \epsilon) \leq K \sqrt{\frac{\epsilon d \log \frac{2}{\epsilon}}{m}}$$

for some constant K [see e.g., 28]. Noting that $\phi_{\mathbb{C}}(m, \epsilon) \leq \omega_{\mathbb{C}}(m, diam(\epsilon; \mathbb{C}))$, we have that $\forall \epsilon > 0$,

$$\tilde{U}_{\mathbb{C}}(m, \epsilon, \delta)$$

$$\leq \tilde{K} \left(K \sqrt{\frac{diam(\tilde{c}\epsilon; \mathbb{C}) d \log \frac{2}{diam(\tilde{c}\epsilon;\mathbb{C})}}{m}} + \sqrt{\frac{s_m(\delta) diam(\tilde{c}\epsilon; \mathbb{C})}{m}} + \frac{s_m(\delta)}{m}\right)$$

$$\leq K' \max \left\{ \sqrt{\frac{\epsilon^{1/\kappa} d \log \frac{2+2\epsilon}{\epsilon}}{m}}, \sqrt{\frac{s_m(\delta) \epsilon^{1/\kappa}}{m}}, \frac{s_m(\delta)}{m}\right\}.$$

If $s_m(\delta)/m \leq 1$, then taking any $\epsilon \geq K''\left(\frac{d\log m+\log(1/\delta)}{m}\right)^{\frac{\kappa}{2\kappa-1}}$, for some appropriate constant $K'' > 0$ (depending on μ and κ), suffices to make this latter quantity $\leq \frac{\epsilon}{16}$. So we must have that

$$
(17) \qquad \min\left\{\tilde{\mathcal{E}}_{\mathbb{C}}(m,\delta), 1\right\} \leq K^{(3)}\left(\frac{d\log m + \log(1/\delta)}{m}\right)^{\frac{\kappa}{2\kappa-1}}
$$

for some constant $K^{(3)}$. Plugging this into the query bound (16), we have that $n \leq$

$$
(18)
$$
$$
\log_2\frac{4m_n^2}{\delta}+4e\theta\left(5+\int_1^{m_n-1}\max\{\mu,1\}(6K^{(3)})^{\frac{1}{\kappa}}\left(\frac{d\log x+\log(1/\delta)}{x}\right)^{\frac{1}{2\kappa-1}}dx\right).
$$

If $\kappa > 1$, (18) is at most $K^{(4)}\theta m_n^{\frac{2\kappa-2}{2\kappa-1}}(d\log m_n + \log(1/\delta))^{\frac{1}{2\kappa-1}}$, for some constant $K^{(4)}$ (depending on κ and μ). This implies

$$
m_n \geq K^{(5)}\left(\frac{n}{\theta(d\log n + \log(1/\delta))^{1/(2\kappa-1)}}\right)^{\frac{2\kappa-1}{2\kappa-2}},
$$

for some constant $K^{(5)}$. Plugging this into (17) completes the proof for this case.

On the other hand, if $\kappa = 1$, (18) is at most $K^{(6)}\theta(d\log m_n+\log(1/\delta))\log m_n$, for some constant $K^{(6)}$ (depending on μ). This implies

$$
m_n \geq exp\left\{K^{(7)}\sqrt{\frac{n}{\theta(d+\log(1/\delta))}}-1\right\},
$$

for some constant $K^{(7)}$. Plugging this into (17) and simplifying the expression with a bit of algebra completes this case.

For the bound under Condition 2, again we have that

$$
\phi_{\mathbb{C}}(m,\epsilon) \leq \omega_{\mathbb{C}}(m, diam(\epsilon;\mathbb{C})) \leq \alpha\cdot\max\left\{\mu^{\frac{1-\rho}{2}}\cdot m^{-1/2}\epsilon^{\frac{1-\rho}{2\kappa}}, m^{-\frac{1}{1+\rho}}\right\}.
$$

From this, it quickly follows that

$$
(19)
$$
$$
\min\left\{\tilde{\mathcal{E}}_{\mathbb{C}}(m,\delta), 1\right\}\leq K^{(8)}\max\left\{m^{-\frac{\kappa}{2\kappa+\rho-1}}, \left(\frac{\log\frac{m}{\delta}}{m}\right)^{\frac{\kappa}{2\kappa-1}}\right\}\leq K^{(8)}\left(\frac{\log\frac{m}{\delta}}{m}\right)^{\frac{\kappa}{2\kappa+\rho-1}}
$$

for some constant $K^{(8)}$ (depending on μ, α, ρ and κ). Plugging this into the query bound (16), we have that

$$
n \leq \log_2\frac{4m_n^2}{\delta} + 4e\theta\left(5+\int_1^{m_n-1}\max\{\mu,1\}(6K^{(8)})^{\frac{1}{\kappa}}\left(\frac{\log\frac{x}{\delta}}{x}\right)^{\frac{1}{2\kappa+\rho-1}}dx\right)
$$
$$
\leq K^{(9)}\theta m_n^{\frac{2\kappa+\rho-2}{2\kappa+\rho-1}}\left(\log\frac{m_n}{\delta}\right)^{\frac{1}{2\kappa+\rho-1}},
$$

for some constant $K^{(9)}$ (depending on κ, μ, α, and ρ). This implies

$$m_n \geq K^{(10)} \left(\frac{n}{\theta \left(\log(n/\delta) \right)^{1/(2\kappa+\rho-1)}} \right)^{\frac{2\kappa+\rho-1}{2\kappa+\rho-2}},$$

for some constant $K^{(10)}$. Plugging this into (19) completes the proof of this case.

□

PROOF OF THEOREM 5 AND THEOREM 6. These theorems now follow from Lemmas 7 and 8. □

B.2. Proofs Relating to Section 5.

LEMMA 9. *For $i \in \mathbb{N}$, let $\delta_i = \delta/(2i^2)$ and $m_{in} = |\mathcal{L}_{in}| + |Q_{in}|$ (for $i > \sqrt{n/2}$, define $\mathcal{L}_{in} = Q_{in} = \emptyset$). For each n, let $\hat{i}_n$ denote the smallest index i satisfying the condition on h_{in} in Step 3 of Algorithm 3. Let $\tau_n = 2^{-n}$ and define*

$$i_n^* = \min \left\{ i \in \mathbb{N} : \forall i' \geq i, \forall j \geq i', \forall h \in \mathbb{C}_{i'}(\tau_n), er_{\mathcal{L}_{jn}}(h) = 0 \right\},$$

and

$$j_n^* = \operatorname*{argmin}_{j \in \mathbb{N}} \nu_j + \hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn}, \delta_j).$$

Then on the event $\bigcap_{i=1}^{\infty} E_{\mathbb{C}_i, \delta_i}$,

$$\forall n \in \mathbb{N}, \max \left\{ i_n^*, \hat{i}_n \right\} \leq j_n^*.$$

◇

PROOF OF LEMMA 9. As before, note that for $\ell \in \mathbb{N} \cup \{0\}$ with $\ell \leq m_{in}$, $\mathcal{L}_{in}^{(\ell)}$ and $Q_{in}^{(\ell)}$ denote the sets $\mathcal{L}$ and Q, respectively, at the conclusion of round ℓ of Algorithm 2, when run with class $\mathbb{C}_i$, label budget $\lfloor n/(2i^2) \rfloor$, confidence parameter δ_i, and threshold as in (11).

Assume the event $\bigcap_{i=1}^{\infty} E_{\mathbb{C}_i, \delta_i}$ occurs. Suppose, for the sake of contradiction, that $j = j_n^* < i_n^*$ for some $n \in \mathbb{N}$. Then there is some $i \geq i_n^* - 1$ such that, for some $\ell < m_{in}$, we have some $h' \in \mathbb{C}_{i_n^*-1}(\tau_n) \cap \mathbb{C}_i[\mathcal{L}_{in}^{(\ell)}]$ with

$$er_\ell(h') - \min_{h \in \mathbb{C}_i} er_\ell(h) \geq er_\ell(h') - \min_{h \in \mathbb{C}_i[\mathcal{L}_{in}^{(\ell)}]} er_\ell(h)$$

$$> 3\hat{\mathcal{E}}_{\mathbb{C}_i}(\mathcal{L}_{in}^{(\ell)} \cup Q_{in}^{(\ell)}, \delta_i; \mathcal{L}_{in}^{(\ell)}) = 3\hat{\mathcal{E}}_{\mathbb{C}_i}(\ell, \delta_i),$$

where the last equality is due to Lemma 6. Lemma 5 implies this will not happen for $i = i_n^* - 1$, so we can assume $i \geq i_n^*$. We therefore have (by Lemma 5) that

$$3\hat{\mathcal{E}}_{\mathbb{C}_i}(\ell, \delta_i) < er_\ell(h') - \min_{h \in \mathbb{C}_i} er_\ell(h) \leq \frac{3}{2} \max \left\{ \tau_n + \nu_{i_n^* - 1} - \nu_i, \hat{\mathcal{E}}_{\mathbb{C}_i}(\ell, \delta_i) \right\}.$$

In particular, this implies that

$$3\hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \leq 3\hat{\mathcal{E}}_{\mathbb{C}_i}(\ell, \delta_i) < \frac{3}{2} \left(\tau_n + \nu_{i_n^* - 1} - \nu_i \right) \leq \frac{3}{2} \left(\tau_n + \nu_j - \nu_i \right).$$

Therefore, (by definition of $j = j_n^*$)

$$\hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn}, \delta_j) + \nu_j \leq \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) + \nu_i \leq \frac{1}{2} \left(\tau_n + \nu_j - \nu_i \right) + \nu_i \leq \frac{\tau_n}{2} + \nu_j.$$

This would imply that $\hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn}, \delta_j) \leq \tau_n/2 < \frac{1}{m_{jn}}$ (due to the second return condition in Algorithm 2), which by definition is not possible, so we have a contradiction. Therefore, we must have that every $j_n^* \geq i_n^*$. In particular, we have that $\forall n \in \mathbb{N}, h_{j_n^* n} \neq \varnothing$.

Now pick an arbitrary $i \in \mathbb{N}$ with $i > j = j_n^*$, and let $h' \in \mathbb{C}_j(\tau_n)$. Then

$er_{\mathcal{L}_{in} \cup Q_{in}}(h_{jn}) - er_{\mathcal{L}_{in} \cup Q_{in}}(h_{in}) = er_{m_{in}}(h_{jn}) - er_{m_{in}}(h_{in})$

$\leq er_{m_{in}}(h_{jn}) - \min_{h \in \mathbb{C}_i} er_{m_{in}}(h)$

$\leq \frac{3}{2} \max \left\{ er(h_{jn}) - \nu_i, \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \right\}$ (by Lemma 5)

$= \frac{3}{2} \max \left\{ er(h_{jn}) - \nu_j + \nu_j - \nu_i, \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \right\}$

$\leq \frac{3}{2} \max \begin{cases} 2(er_{m_{jn}}(h_{jn}) - er_{m_{jn}}(h') + \tau_n) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn}, \delta_j) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \end{cases}$ (by Lemma 5)

$= \frac{3}{2} \max \begin{cases} \hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn}, \delta_j) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \end{cases}$ (since $j \geq i_n^*$)

$= \frac{3}{2} \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i)$ (by definition of j_t^*)

$= \frac{3}{2} \hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}_{in} \cup Q_{in}, \delta_i; \mathcal{L}_{in})$ (by Lemma 6).

$\square$

LEMMA 10. *On the event* $\bigcap\limits_{i=1}^{\infty} E_{\mathbb{C}_i, \delta_i}, \forall n \in \mathbb{N}$,

$$er(h_{\hat{i}_n n}) - \nu_\infty \leq 3 \min_{i \in \mathbb{N}} \left(\nu_i - \nu_\infty + \tilde{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \right).$$

$\diamond$

PROOF OF LEMMA 10. Let $h'_n \in \mathbb{C}_{j^*_n}(\tau_n)$ for $\tau_n \in (0, 2^{-n})$, $n \in \mathbb{N}$. Suppose the event $\bigcap_{i=1}^{\infty} E_{\mathbb{C}_i, \delta_i}$ occurs.

$$er(h_{\hat{i}_n n}) - \nu_\infty$$
$$= \nu_{j^*_n} - \nu_\infty + er(h_{\hat{i}_n n}) - \nu_{j^*_n}$$
$$\leq \nu_{j^*_n} - \nu_\infty + \max \begin{cases} 2(er_{m_{j^*_n n}}(h_{\hat{i}_n n}) - er_{m_{j^*_n n}}(h'_n) + \tau_n) \\ \hat{\mathcal{E}}_{\mathbb{C}_{j^*_n}}(m_{j^*_n n}, \delta_{j^*_n}) \end{cases}$$
$$\leq \nu_{j^*_n} - \nu_\infty + \max \begin{cases} 2(er_{\mathcal{L}_{j^*_n n} \cup Q_{j^*_n n}}(h_{\hat{i}_n n}) - er_{\mathcal{L}_{j^*_n n} \cup Q_{j^*_n n}}(h_{j^*_n n}) + \tau_n) \\ \hat{\mathcal{E}}_{\mathbb{C}_{j^*_n}}(m_{j^*_n n}, \delta_{j^*_n}) \end{cases}$$

The first inequality follows from Lemma 5. The second inequality is due to Lemma 9 (i.e., $j^*_n \geq \max\{i^*_n, \hat{i}_n\}$). In this last line, we can let $\tau_n \to 0$, and use the definition of $\hat{i}_n$ combined with the fact that $\hat{i}_n \leq j^*_n$, to show that it is at most

$$\nu_{j^*_n} - \nu_\infty + \max \left\{ 2\left(\frac{3}{2} \hat{\mathcal{E}}_{\mathbb{C}_{j^*_n}}(\mathcal{L}_{j^*_n n} \cup Q_{j^*_n n}, \delta_{j^*_n}; \mathcal{L}_{j^*_n n}) \right), \hat{\mathcal{E}}_{\mathbb{C}_{j^*_n}}(m_{j^*_n n}, \delta_{j^*_n}) \right\}$$
$$= \nu_{j^*_n} - \nu_\infty + 3\hat{\mathcal{E}}_{\mathbb{C}_{j^*_n}}(m_{j^*_n n}, \delta_{j^*_n}) \qquad \text{(by Lemma 6)}$$
$$\leq 3 \min_i \left(\nu_i - \nu_\infty + \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \right) \qquad \text{(by definition of } j^*_n)$$
$$\leq 3 \min_i \left(\nu_i - \nu_\infty + \tilde{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i) \right) \qquad \text{(by Lemma 5).}$$

$\square$

PROOF OF THEOREM 7 AND THEOREM 8. These theorems now follow from Lemmas 10 and 8. That is, Lemma 10 gives a bound in terms of the $\tilde{\mathcal{E}}$ quantities, holding on event $\bigcap_{i=1}^{\infty} E_{\mathbb{C}_i, \delta_i}$, and Lemma 8 bounds these $\tilde{\mathcal{E}}$ quantities as desired, on event $\bigcap_{i=1}^{\infty} H_{\mathbb{C}_i, \delta_i} \cap E_{\mathbb{C}_i, \delta_i}$. Noting that, by a union bound,

$$\mathbb{P}\left[\bigcap_{i=1}^{\infty} H_{\mathbb{C}_i, \delta_i} \cap E_{\mathbb{C}_i, \delta_i} \right] \geq 1 - \sum_{i=1}^{\infty} \delta_i \geq 1 - \delta$$

completes the proof. $\square$

DEFINITION 5. *Define* $\mathring{c} = \tilde{c} + 1$, $\mathring{D}(\epsilon) = \lim_{j \to \infty} diam(\epsilon; \mathbb{C}_j)$, *and*

$$\mathring{U}_{\mathbb{C}_i}(m, \epsilon, \delta_i) = \tilde{K} \left(\omega_{\mathbb{C}_i}(m, \mathring{D}(\mathring{c}\epsilon)) + \sqrt{\frac{s_m(\delta_i)\mathring{D}(\mathring{c}\epsilon)}{m}} + \frac{s_m(\delta_i)}{m} \right)$$

and $$\mathring{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i) = \inf \left\{ \epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \mathring{U}_{\mathbb{C}_i}(m, 2^j, \delta_i) \leq 2^{j-4} \right\}.$$

$\diamond$

LEMMA 11.   *For any $m, i \in \mathbb{N}$,*

$$\tilde{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i) \leq \max\left\{\mathring{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i), \nu_i - \nu_\infty\right\}.$$

$\diamond$

PROOF OF LEMMA 11. For $\epsilon > \nu_i - \nu_\infty$,

$$\tilde{U}_{\mathbb{C}_i}(m, \epsilon, \delta_i) = \tilde{K}\left(\phi_{\mathbb{C}_i}(m, \tilde{c}\epsilon) + \sqrt{\frac{s_m(\delta_i)diam(\tilde{c}\epsilon; \mathbb{C}_i)}{m}} + \frac{s_m(\delta_i)}{m}\right)$$

$$\leq \tilde{K}\left(\omega_{\mathbb{C}_i}(m, diam(\tilde{c}\epsilon; \mathbb{C}_i)) + \sqrt{\frac{s_m(\delta_i)diam(\tilde{c}\epsilon; \mathbb{C}_i)}{m}} + \frac{s_m(\delta_i)}{m}\right).$$

But $diam(\tilde{c}\epsilon; \mathbb{C}_i) \leq \mathring{D}(\tilde{c}\epsilon + (\nu_i - \nu_\infty)) \leq \mathring{D}(\mathring{c}\epsilon)$, so the above line is at most

$$\tilde{K}\left(\omega_{\mathbb{C}_i}(m, \mathring{D}(\mathring{c}\epsilon)) + \sqrt{\frac{s_m(\delta_i)\mathring{D}(\mathring{c}\epsilon)}{m}} + \frac{s_m(\delta_i)}{m}\right) = \mathring{U}_{\mathbb{C}_i}(m, \epsilon, \delta_i).$$

In particular, this implies that

$$\begin{aligned}\tilde{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i) =\quad & \inf\left\{\epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \tilde{U}_{\mathbb{C}_i}(m, 2^j, \delta_i) \leq 2^{j-4}\right\} \\ \leq\quad & \inf\left\{\epsilon > (\nu_i - \nu_\infty) : \forall j \in \mathbb{Z}_\epsilon, \tilde{U}_{\mathbb{C}_i}(m, 2^j, \delta_i) \leq 2^{j-4}\right\} \\ \leq\quad & \inf\left\{\epsilon > (\nu_i - \nu_\infty) : \forall j \in \mathbb{Z}_\epsilon, \mathring{U}_{\mathbb{C}_i}(m, 2^j, \delta_i) \leq 2^{j-4}\right\} \\ =\quad & \max\left\{\inf\left\{\epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \mathring{U}_{\mathbb{C}_i}(m, 2^j, \delta_i) \leq 2^{j-4}\right\}, (\nu_i - \nu_\infty)\right\} \\ =\quad & \max\left\{\mathring{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i), \nu_i - \nu_\infty\right\}. \qquad\qquad \square\end{aligned}$$

PROOF OF THEOREM 9. By the same argument that gave us (17), we have that

$$\min\left\{\mathring{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i), 1\right\} \leq K_1 \frac{d_i \log m + \log(i/\delta)}{m},$$

for some constant K_1 (depending on μ).

Now assume the event $\bigcap_{i=1}^\infty H_{\mathbb{C}_i, \delta_i} \cap E_{\mathbb{C}_i, \delta_i}$ occurs. In particular, Lemmas 10 and 11 imply that for some constant K_2, $\forall n \in \mathbb{N}$,

$$er(\hat{h}_n) - \nu^* \leq \min\left\{1, 3 \min_{i \in \mathbb{N}}\left(2(\nu_i - \nu^*) + \mathring{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i)\right)\right\}$$

$$(20) \qquad\qquad \leq K_2 \min_{i \in \mathbb{N}}\left((\nu_i - \nu^*) + \min\left\{1, \frac{d_i \log m_{in} + \log(i/\delta)}{m_{in}}\right\}\right).$$

Take any $i \in \mathbb{N}$. The label request bound of Lemma 7, along with Lemma 11, implies that

$$\lfloor n/(2i^2) \rfloor$$

$$\leq \log_2 \frac{8m_{in}^2 i^2}{\delta} + K_3 \theta_i \left(5 + \int_1^{m_{in}} \max \left\{ \nu_i - \nu^*, \frac{d_i \log x + \log(i/\delta)}{x} \right\} dx \right)$$

$$\leq K_4 \theta_i \max \left\{ (\nu_i - \nu^*) m_{in}, (d_i \log m_{in} + \log(i/\delta)) \log m_{in} \right\}.$$

Letting $\gamma_i(n) = \sqrt{\frac{n}{i^2 \theta_i (d_i + \log(i/\delta))}}$, this implies that

$$\min \left\{ 1, \frac{d_i \log m_{in} + \log(i/\delta)}{m_{in}} \right\}$$

$$\leq \min \left\{ 1, K_5 \left((\nu_i - \nu^*) + \left(d_i(\gamma_i(n) + \gamma_i(n)^2) + \log \frac{i}{\delta} \right) exp \left\{ -c_2 \gamma_i(n) \right\} \right) \right\}$$

$$\leq K_6 \left((\nu_i - \nu^*) + \left(d_i + \log \frac{i}{\delta} \right) exp \left\{ -c_3 \gamma_i(n) \right\} \right).$$

Combined with (20), this establishes the result. $\qquad\qquad\qquad\qquad\square$

APPENDIX C: ESTIMATORS IN ALGORITHM 1

As mentioned earlier, we can replace $\mathbb{P}(R)$ and $\mathbb{P}(DIS(V))$ in Algorithm 1 with estimators based only on *unlabeled* data points, while preserving the rates in Theorem 4 up to constant factors. Here we briefly sketch the modifications to the proof necessary to compensate for this additional estimation.

The estimators can be defined in a variety of ways. To be concrete here, we define them based on Hoeffding's inequality, letting $M_m = \frac{512(m+1)^6}{\delta^2} \ln \frac{8(m+1)^2}{\delta}$, $\Gamma_m = \frac{\delta}{32(m+1)^3}$, $\hat{\mathbb{P}}_m(DIS(V)) = \Gamma_m + M_m^{-1} \sum_{i=1}^{M_m} \mathbb{1}_{DIS(V)} \left(X'_{i,m} \right)$, and $\hat{\mathbb{P}}_m(R) = \Gamma_m + M_m^{-1} \sum_{i=1}^{M_m} \mathbb{1}_R \left(X'_{i,m} \right)$, where the $X'_{i,m}$ are distributed i.i.d. according to $\mathcal{D}_X$, independent from the data sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ (perhaps set aside in a preprocessing step). Replacing each "$\mathbb{P}$" in Algorithm 1 with "$\hat{\mathbb{P}}_m$" (for the current value of m in the algorithm), the analysis above can be adapted to compensate as follows.

First, note that by Hoeffding's inequality and a union bound, with probability $1 - \delta$, every $\hat{\mathbb{P}}_m(DIS(V))$ and $\hat{\mathbb{P}}_m(R)$ calculated in the algorithm satisfies $\hat{\mathbb{P}}_m(DIS(V)) - 2\Gamma_m \leq \mathbb{P}(DIS(V)) \leq \hat{\mathbb{P}}_m(DIS(V))$ and $\hat{\mathbb{P}}_m(R) - 2\Gamma_m \leq \mathbb{P}(R) \leq \hat{\mathbb{P}}_m(R)$; supposing this $1 - \delta$ probability event occurs, along with the $1 - \delta$ probability event analogous to that in the original proof above (i.e., that the UB and LB evaluations are valid), the β_t values remain valid bounds on the achieved excess error rate. Also, at any time with $\mathbb{P}(R) \leq 1/2$, consider $m' \leq$

$-\ln(1 - \delta/(2(m + 1)^2))/(2\mathbb{P}(R)) \leq \ln(1 - \delta/(2(m + 1)^2))/\ln(1 - \mathbb{P}(R))$; we have $(1-\mathbb{P}(R))^{m'} \geq 1-\delta/(2(m+1)^2)$, so that with probability $1-\delta/(2(m+1)^2)$ the next m' data points are not in R. Thus, by a union bound, with probability $1 - \delta$, at all times in the algorithm we have $\min\{m' > m : X_{m'} \in R\} - m > -\ln(1 - \delta/(2(m + 1)^2))/(2\mathbb{P}(R))$. Suppose this event also holds. In particular, this implies the invariant that when calculating $\hat{\mathbb{P}}_m$ in Steps 2 and 9 for $m > 0$, $\Gamma_m = \frac{\delta}{32(m+1)^3} \leq \frac{-\ln(1-\delta/(2(m+1)^2))}{16(m+1)} \leq \mathbb{P}(R)/8$.

Now replace the "2" in (4) with "4." By the same reasoning as before, if $V^{(\theta)} = \emptyset$, we have $\mathbb{P}(DIS(V)) \leq \mathbb{P}(R)/4$, so that $\hat{\mathbb{P}}_m(DIS(V)) \leq \mathbb{P}(DIS(V)) + 2\Gamma_m \leq \mathbb{P}(DIS(V)) + \mathbb{P}(R)/4 \leq \mathbb{P}(R)/2 \leq \hat{\mathbb{P}}_m(R)/2$, which satisfies the condition in Step 2 on the next round. Propagating this change in $V^{(\theta)}$, (5) becomes $\mathbb{P}(R)^{\kappa-1}(4\mu\theta)^{-\kappa} < 4G(|Q| - 1, \delta/n)$. Also, (6) becomes $\frac{\epsilon}{2G(|Q|-1,\delta/n)} < \hat{\mathbb{P}}_m(R) \leq \mathbb{P}(R) + 2\Gamma_m \leq 2\mathbb{P}(R)$, so that we may simply replace each "2" in (7) by a "4," and consequently the same is true of (8); in (9) and (10), this changes the "$\frac{2}{\epsilon}$" to "$\frac{4}{\epsilon}$" and the "6" to "12." Finally, in addition to these adjustments to (10), the $\log_2 \frac{2}{\epsilon}$ factor in (10) becomes $\log_{8/5} \frac{2}{\epsilon}$, and is explained as follows. When the condition in Step 2 is satisfied for some m, $\mathbb{P}(DIS(V)) \leq \hat{\mathbb{P}}_m(DIS(V)) \leq (1/2)\hat{\mathbb{P}}_m(R) \leq (1/2)\mathbb{P}(R) + \Gamma_m \leq (5/8)\mathbb{P}(R)$, and therefore $\mathbb{P}(R)$ is reduced at least by a factor of $5/8$ every time we reach Step 3. In particular, after the algorithm satisfies the condition in Step 2 at most $\left\lceil \log_{8/5} \frac{5}{4\epsilon} \right\rceil \leq \log_{8/5} \frac{2}{\epsilon}$ times, we are guaranteed some t (and corresponding R) satisfies $\beta_t \leq \hat{\mathbb{P}}_m(R) \leq \mathbb{P}(R) + 2\Gamma_m \leq (5/4)\mathbb{P}(R) \leq \epsilon$. This guarantees the stated excess error bound from Theorem 4 (with the slightly adjusted constants indicated above), holding with probability at least $1 - 3\delta$.

APPENDIX D: A GENERAL MINIMAX LOWER BOUND

In this appendix, we prove a general minimax lower bound for nontrivial $\mathbb{C}$ under Condition 1, matching the result of Castro and Nowak [13] for threshold classifiers (though stated in a slightly different form). The proof is essentially similar to that of Kääriäinen [22], with modifications suited to these noise conditions.

For any hypothesis class $\mathbb{C}$, and any $\mu, \kappa \in [1, \infty)$, let $\mathcal{T}(\mathbb{C}, \mu, \kappa)$ denote the set of distributions $\mathcal{D}_{XY}$ satisfying Condition 1 for the hypothesis class $\mathbb{C}$ and the specified parameters μ and κ. Also, let $\mathbb{A}$ denote the set of all valid active learning algorithms (of the kind studied above, taking as input a budget n and a confidence parameter δ, and returning a classifier after at most n label requests).

THEOREM 10. *For any hypothesis class $\mathbb{C}$ with $|\mathbb{C}| \geq 3$, and any $\mu \in [2, \infty)$ and $\kappa \in (1, \infty)$, there is a κ-dependent constant $c \in (0, \infty)$ such that, for any $\delta \in (0, 1/16)$ and any integer $n \in \mathbb{N}$,*

$$\inf_{\mathcal{A} \in \mathbb{A}} \sup_{\mathcal{D}_{XY} \in \mathcal{T}(\mathbb{C}, \mu, \kappa)} \mathbb{P} \left(er \left(\mathcal{A}(n, \delta) \right) - \nu \geq c n^{-\frac{\kappa}{2\kappa - 2}} \right) > \delta.$$

$\diamond$

PROOF. We proceed by reduction from hypothesis testing. Specifically, consider two values: $p_0 = \frac{1}{2} \left(1 - \epsilon^{\frac{\kappa - 1}{\kappa}} \right)$ and $p_1 = \frac{1}{2} \left(1 + \epsilon^{\frac{\kappa - 1}{\kappa}} \right)$, where we will specify a value for $\epsilon \in (0, 1)$ later. Consider the problem of constructing an estimator $\hat{i}_n(b_1, \ldots, b_n) \in \{0, 1\}$, where each $b_i \in \{0, 1\}$. It is known [e.g., 10, 34] that for any $\delta \in (0, 1/16)$, if

(21)
$$n < \frac{(1 - 8\delta) \ln \frac{1}{8\delta}}{8 KL (p_0 \| p_1)}$$

(for $KL(p_0 \| p_1) = p_0 \ln \frac{p_0}{p_1} + (1 - p_0) \ln \frac{1 - p_0}{1 - p_1}$), then for any such $\hat{i}_n$ estimator (possibly including additional independent internal randomness), there is some $i \in \{0, 1\}$ for which, if $B_1^{(i)}, \ldots, B_n^{(i)}$ are i.i.d. Bernoulli(p_i) random variables, then

(22)
$$\mathbb{P} \left(\hat{i}_n(B_1^{(i)}, \ldots, B_n^{(i)}) \neq i \right) > \delta.$$

Now returning to the active learning problem, we design the distribution $\mathcal{D}_X$ as follows. Since $|\mathbb{C}| \geq 3$, there exist $h_0, h_1 \in \mathbb{C}$ and $x, x' \in \mathcal{X}$ such that $h_0(x) \neq h_1(x)$ and $h_0(x') = h_1(x')$. Suppose $0 < \epsilon < 1 - 2^{-\frac{1}{\kappa}}$; we will specify a specific value for ϵ later. Assign $\mathcal{D}_X(\{x\}) = \epsilon^{\frac{1}{\kappa}}$, and $\mathcal{D}_X(\{x'\}) = 1 - \epsilon^{\frac{1}{\kappa}}$. Note that $\mathcal{D}_X(\{x'\}) > 1 - \left(1 - 2^{-\frac{1}{\kappa}} \right)^{\frac{1}{\kappa}} \geq 1 - 2^{-\frac{1}{\kappa}} > \epsilon$.

Now fix $\delta \in (0, 1/16)$, and suppose we have some active learning algorithm $\mathcal{A}$. We will construct an estimator $\hat{i}_n$ based on $\mathcal{A}$. Specifically, we start by randomly sampling a sequence $X_1, X_2, \ldots$ i.i.d. according to $\mathcal{D}_X$, to serve as the data sequence for $\mathcal{A}$. Now we run $\mathcal{A}(n, \delta)$. Every X_j is equal either x or x'. For the tth label request made by the algorithm for some $X_j = x$, we (in this case, playing the role of the oracle) return as the label of X_j the value $h_{b_t}(x)$. The other label requests are for the label of some $X_j = x'$, and we return $h_0(x')$ for these. In the end, $\mathcal{A}$ returns a classifier $\hat{h}$; based on this, define $\hat{i}_n(b_1, \ldots, b_n) = \mathbb{1}[\hat{h}(x) = h_1(x)]$.

This simulated behavior for $\mathcal{A}$ to calculate $\hat{i}_n(B_1^{(i)}, \ldots, B_n^{(i)})$, where $B_1^{(i)}, \ldots, B_n^{(i)}$ are i.i.d. Bernoulli(p_i), is distributionally equivalent to running $\mathcal{A}(n, \delta)$ under a distribution $\mathcal{D}_{XY} = \mathcal{D}_i$, where $\mathcal{D}_i$ has marginal $\mathcal{D}_X$ on $\mathcal{X}$, and is otherwise defined by the following property. For $(X, Y) \sim \mathcal{D}_i$, $\mathbb{P}(Y = h_i(x) | X = x) =$

$\frac{1}{2}\left(1 + \epsilon^{\frac{\kappa-1}{\kappa}}\right)$, and $\mathbb{P}(Y = h_i(x')|X = x') = 1$. Thus, for any h' with $h'(x) \neq h_i(x)$ and $h'(x') = h_i(x')$, we have (for $\mathcal{D}_{XY} = \mathcal{D}_i$)

$$
(23) \qquad er(h') - er(h_i) = (er(h'|\{x\}) - er(h_i|\{x\}))\epsilon^{\frac{1}{\kappa}}
$$

$$
= \left(\frac{1}{2}\left(1 + \epsilon^{\frac{\kappa-1}{\kappa}}\right) - \frac{1}{2}\left(1 - \epsilon^{\frac{\kappa-1}{\kappa}}\right)\right)\epsilon^{\frac{1}{\kappa}} = \epsilon;
$$

in particular, this is true for $h' = h_{1-i}$. Since any h' with $h'(x') \neq h_i(x')$ has $er(h') - er(h_i) \geq 1 - \epsilon^{\frac{1}{\kappa}} > \epsilon$, we know that for $\epsilon \leq \gamma < 1 - \epsilon^{\frac{1}{\kappa}}$, $diam(\gamma; \mathbb{C}) = diam(\epsilon; \mathbb{C}) = \epsilon^{\frac{1}{\kappa}} \leq \mu\gamma^{\frac{1}{\kappa}}$. Furthermore, combining this with (23), we have that any h' with either $h'(x) \neq h_i(x)$ or $h'(x') \neq h_i(x')$ has $er(h') - er(h_i) \geq \epsilon$. Thus, for $0 < \gamma < \epsilon$, $diam(\gamma; \mathbb{C}) = 0 < \mu\gamma^{\frac{1}{\kappa}}$. Finally, for $1 - \epsilon^{\frac{1}{\kappa}} \leq \gamma \leq 1$,

$$
diam(\gamma; \mathbb{C}) \leq 1 \leq \left(\frac{1}{1 - \epsilon^{\frac{1}{\kappa}}}\right)^{\frac{1}{\kappa}}\gamma^{\frac{1}{\kappa}} < \left(\frac{1}{1 - \left(1 - 2^{-\frac{1}{\kappa}}\right)^{\frac{1}{\kappa}}}\right)^{\frac{1}{\kappa}}\gamma^{\frac{1}{\kappa}} \leq 2\gamma^{\frac{1}{\kappa}} \leq \mu\gamma^{\frac{1}{\kappa}}.
$$

We therefore have that $\mathcal{D}_i \in \mathcal{T}(\mathbb{C}, \mu, \kappa)$.

Now suppose $\hat{h}'$ is the returned value from running $\mathcal{A}(n, \delta)$ with distribution $\mathcal{D}_{XY}$. As mentioned, for any given $i \in \{0, 1\}$, if $\mathcal{D}_{XY} = \mathcal{D}_i$, then $\mathbb{1}[\hat{h}'(x) = h_1(x)]$ is distributionally equivalent to $\hat{i}_n(B_1^{(i)}, \ldots, B_n^{(i)})$ for $B_1^{(i)}, \ldots, B_n^{(i)}$ i.i.d. Bernoulli(p_i). The reasoning above also indicates that for $\mathcal{D}_{XY} = \mathcal{D}_i$,

$$
\mathbb{P}\left(er\left(\hat{h}'\right) - er(h_i) \geq \epsilon\right) \geq \mathbb{P}\left(\hat{h}'(x) \neq h_i(x)\right)
$$

$$
= \mathbb{P}\left(\mathbb{1}[\hat{h}'(x) = h_1(x)] \neq i\right) = \mathbb{P}\left(\hat{i}_n(B_1^{(i)}, \ldots, B_n^{(i)}) \neq i\right).
$$

Therefore, by the aforementioned known results on hypothesis testing for Bernoulli means, for n as in (21),

$$
\sup_{\mathcal{D}_{XY} \in \{\mathcal{D}_0, \mathcal{D}_1\}} \mathbb{P}\left(er\left(\hat{h}'\right) - \nu \geq \epsilon\right) \geq \sup_{i \in \{0,1\}} \mathbb{P}\left(\hat{i}_n(B_1^{(i)}, \ldots, B_n^{(i)}) \neq i\right) > \delta.
$$

Studying (21), and noting that $\gamma \in (0, 1/2) \implies KL\left(\frac{1}{2}(1 - \gamma)\|\frac{1}{2}(1 + \gamma)\right) \leq 2\ln(3)\gamma^2$, we see that taking

$$
\epsilon = \left(1 - 2^{-\frac{1}{\kappa}}\right)(36n)^{-\frac{\kappa}{2\kappa-2}} < 1 - 2^{-\frac{1}{\kappa}}
$$

satisfies

$$
n \leq \frac{1}{36}\epsilon^{\frac{2-2\kappa}{\kappa}}\ln\frac{1}{8\delta} < \frac{(1 - 8\delta)\ln\frac{1}{8\delta}}{8KL(p_0\|p_1)}.
$$

$\square$

Although this is stated for the worst case over the distribution $\mathcal{D}_{XY}$, including the worst case $\mathcal{D}_X$, we can extend the proof to most nontrivial fixed $\mathcal{D}_X$, only maximizing over $\mathcal{D}_{XY}$ subject to the constraint that it has marginal $\mathcal{D}_X$; specifically, it suffices to have $\mathcal{D}_X$ such that $\exists h_1, h_2 \in \mathbb{C}$ with $\mathbb{P}(h_1(X) \neq h_2(X)) \propto n^{-\frac{1}{2\kappa-2}}$ [see 20, 22]. Also, the same proof technique can be used to show an analogous lower bound on the minimax rates for the *expected* excess error rate, as originally studied by Castro and Nowak [13]. In fact, in that case (21) and (22) clearly imply a minimax lower bound $\exp\{-\Omega(n)\}$ for $\kappa = 1$; for instance, letting $\mathcal{D}_X$ be concentrated entirely on a single point x, using a small constant label noise rate, and setting $\delta = \exp\{-\Omega(n)\}$, for sufficiently large n it is necessary to identify the optimal label for x with probability $1 - \delta$ to guarantee expected excess error $\propto \delta$. Finally, it is interesting to note that the same setup used in the above proof can be used to show a general minimax lower bound $\propto m^{-\frac{\kappa}{2\kappa-1}}$ for *passive learning* from m i.i.d. labeled data points; the proof is identical to that above, except to note that we expect only $\mathcal{D}_X(\{x\})m = \epsilon^{\frac{1}{\kappa}}m$ samples among $\{X_1, \ldots, X_m\}$ will equal x, so that even passive algorithms taking $m \propto \epsilon^{\frac{1-2\kappa}{\kappa}}$ i.i.d. samples as input can be converted into $\hat{i}_n$ estimators based on only $n \propto \epsilon^{\frac{2-2\kappa}{\kappa}}$ Bernoulli samples (since only the X_j equal to x require a sample $B_t^{(i)}$ to generate a label).

REFERENCES

[1] K. S. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *The Annals of Probability*, 12:1041–1067, 1984. 4.4

[2] K. S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer II*, pages 475–493, 1985. 2.2

[3] K. S. Alexander. Sample moduli for set-indexed gaussian processes. *The Annals of Probability*, 14:598–611, 1986. 2.2

[4] 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. 2.2

[5] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 4.1

[6] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006. 1, 2.2, 3.1, 4, 4.4

[7] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007. 1, 2.1, 4.3

[8] 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. 2.2, A.1

[9] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 1

[10] Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, pages 335–344, 2003. D

[11] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *International Conference on Machine Learning*, 2009. 2.2, 2.2

[12] 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. 2

[13] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 1, 2.1, 4.3, 4.4, 4.5, 4.5, D, D

[14] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994. 2.2, 3

[15] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007. 1, 2.2, 2.2, 3.2, 4, 4.2, 3

[16] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996. 2, 3.2, 5

[17] E. Friedman. Active learning for smooth problems. In *Proceedings of the 22nd Conference on Learning Theory*, 2009. 2.2

[18] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 2.2, 4.5

[19] E. Giné, V. Koltchinskii, and J. Wellner. Ratio limit theorems for empirical processes. In *Stochastic Inequalities*, pages 249–278. Birkhäuser, 2003. 4.5

[20] S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007. D

[21] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, 2007. 1, 2.2, 2.2, 1, 2.2, 4.1, 4.2, 2, 4.4

[22] M. Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, 2006. 4.3, D, D

[23] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994. 1

[24] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 2, 2.1, 2.1, 4.5, 4.5, 4.5, 5, A, A, 5, B

[25] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems: Lecture notes. Technical report, Ecole d'ete de Probabilités de Saint-Flour, 2008. 4.5

[26] Y. Li and P. M. Long. Learnability and the doubling dimension. In *Advances in Neural Information Processing*, 2007. 2.2

[27] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27: 1808–1829, 1999. 1, 2.1, 2.1, 4.4

[28] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006. 2, 2.1, 2.1, B.1

[29] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 1, 2.1, 2.1, 4.3, 4.4, 4.5, 4.5, 5

[30] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 2.2, 4.5

[31] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982. 2, 2.2, 3.1, 5

[32] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998. 2, 3.1, 3.2

[33] 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. 2

[34] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. D

[35] L. Wang. Sufficient conditions for agnostic active learnable. In *Advances in Neural Information Processing Systems 22*, 2009. 2.2, 4.5

STEVE HANNEKE
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVE.
PITTSBURGH, PA 15213 USA
E-MAIL: shanneke@stat.cmu.edu