Adaptive Rates of Convergence in Active Learning

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

Abstract

We study the rates of convergence in classification error achievable by active learning in the presence of label noise. Additionally, we study the more general problem of 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 is a powerful supervised learning method capable of producing more accurate classifiers while using a smaller number of labeled examples than traditional (passive) learning techniques. In active learning, a learning algorithm is given access to a large pool of unlabeled examples, and is allowed to interactively request the label of any particular examples from that pool. The objective is to learn a function that accurately predicts the labels of new examples, while minimizing the number of label requests. This contrasts with passive learning, where the labeled examples are sampled at random. In comparison, by more carefully selecting which examples 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 examples are available in abundance, but label information comes only through significant effort or cost.

There have recently been a series of exciting advances on the topic of active learning with arbitrary classification noise (the so-called *agnostic* PAC model), resulting in several new algorithms capable of achieving improved convergence rates compared to passive learning under certain conditions. The first, proposed by [BBL06] was the A^2 (agnostic active) algorithm, which is provably never significantly worse than passive learning by empirical risk minimization. This algorithm was later analyzed in more detail in [Han07], 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 [DHM07], 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 [CN06] 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. That work 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. In a personal communication, John Langford and Rui Castro claimed A^2 achieves these improvements for the special case of threshold classifiers. 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 [MT99, Tsy04]; in particular, we find that these rates are strictly superior to the known rates for passive learning, when the disagreement coefficient is small. We also study a novel modification of the algorithm of [DHM07], 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 each have finite capacity, 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* $\mathcal{X}$, and some fixed distribution $\mathcal{D}_{XY}$ over $\mathcal{X} \times \{-1, 1\}$, with marginal $\mathcal{D}_X$ over $\mathcal{X}$. There is some i.i.d. sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ sampled according to $\mathcal{D}_{XY}$. However, the learning algorithm is only permitted to observe the X_i values (unlabeled examples), and must request the Y_i values one at a time, interactively. That is, the algorithm picks some index i to observe the Y_i value, then after observing it, picks another index i' to request 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 this sequence of examples as being 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 $\delta \in (0, 1)$. The actual number of (unlabeled) examples the algorithm uses will be made clear in the proofs (it is typically close to the passive learning sample complexity corresponding to the stated error guarantee).

A *hypothesis class* $\mathbb{C}$ is simply a set of measurable classifiers $h : \mathcal{X} \to \{-1, 1\}$. We will denote by d the VC dimension of $\mathbb{C}$ [Vap82]. For any measurable $h : \mathcal{X} \to \{-1, 1\}$ and distribution $\mathcal{D}$, define $er_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\{h(X) \neq Y\}$, the *error rate* of h; when $\mathcal{D} = \mathcal{D}_{XY}$, we abbreviate this as $er(h)$. 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 $\mathbb{C}$. Additionally, define the *diameter* of $V \subseteq \mathbb{C}$ 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* as $diam(\epsilon; \mathbb{C}) = diam(\{h \in \mathbb{C} : er(h) - \inf_{h'\in\mathbb{C}} er(h') \leq \epsilon\})$. 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 $\nu^* = er(h^*)$. h^* is called the *Bayes optimal classifier*, and ν^* is the *Bayes error rate*. For a classifier h, and a sequence of labeled examples $Q = \{(X'_1, Y'_1), (X'_2, Y'_2), \ldots, (X'_m, Y'_m)\}$, let $er_Q(h) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}[h(X'_i) \neq Y'_i]$ denote the *empirical error rate* on Q. For the *true* labeled sequence, $\mathcal{Z}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, we abbreviate this by $er_m(h) = er_{\mathcal{Z}_m}(h)$, the true empirical error on the first m examples.

2.1 Tsybakov's Noise Conditions

Here we describe a particular parameterization of noise distributions, relative to a hypothesis class, known as Tsybakov's margin conditions [MT99,Tsy04]. 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., [MT99, Tsy04, Kol06, MN06]).

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

For example, this is satisfied when $h^* \in \mathbb{C}$, and $\exists\mu' > 0, \kappa \geq 1$ s.t. $\forall h \in \mathbb{C}, er(h) - \nu \geq \mu'\mathbb{P}\{h(X) \neq h^*(X)\}^\kappa$, or $\exists\alpha, \mu' > 0$ s.t. $\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq \mu' t^\alpha$, for $t \in (0, 1/2)$ [MT99, Tsy04, Kol06]. 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 this condition can often be interpreted in terms of the relation between magnitude of noise, density, and distance to the decision boundary; that is, in practice the amount of noise in an example's label is often inversely related to the distance from the decision boundary, and the value of κ is essentially determined by how quickly $\eta(x)$ changes as x approaches the decision boundary, relative to how dense the distribution is in that region. See [MT99, Tsy04, Kol06, CN06, MN06] for further interpretations of this margin condition.

It is known that when these conditions are 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

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

where c is a (κ and μ -dependent) constant [Kol06]. Furthermore, for some hypothesis classes, this is known to be a tight bound (up to the log factor) on the minimax 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 [CN06, Tsy04].

2.2 Disagreement Coefficient

The disagreement coefficient, introduced in [Han07], 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 [CAL94, BBL06, DHM07]. 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 $h : \mathcal{X} \to \{-1, 1\}$, 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*[1]

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

We further define the disagreement coefficient for $\mathbb{C}$ with respect to $\mathcal{D}_{XY}$ as $\theta = \liminf_{k\to\infty} \theta_{h^{[k]}}$, where $\{h^{[k]}\}$ is any sequence of $h^{[k]} \in \mathbb{C}$ with $er(h^{[k]})$ monotonically decreasing to ν. ◇

The r_0 in the definition can either be defined as 0, giving a coarse analysis, or for a more subtle analysis we can take it to be a function of n, the number of labels. For our present purposes, we will generally take $r_0 = 0$; a more refined analysis

[1] Throughout this paper, we will let $\mathbb{E}$ and $\mathbb{P}$ (and indeed *any* reference to "probability") refer to the *outer* expectation and measure [vdVW96], so that quantities such as $\mathbb{P}(DIS(B(h, r)))$ are well defined, even if $DIS(B(h, r))$ is not measurable.

with r_0 a function of n will appear in an extended version of this paper.

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 some type of geometric representation.

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 [CAL94] 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 examples in sequence, and for each one it determines whether there exists a classifier in $\mathbb{C}$ consistent with all previously observed labels that labels this new example $+1$ *and* one that labels this example -1; 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 an example it can derive from known information, but otherwise makes no effort to search for informative examples. 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 example 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.

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 examples 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.

The first algorithm, originally proposed by [BBL06], 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. A version of the algorithm is described below.

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

```
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 $h \in V$
5.     $m \leftarrow \min\{m' > m : X_{m'} \in R\}$
6.     Request $Y_m$ and let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$
7.     $V \leftarrow \{h \in V : LB(h, Q, \delta/n) \leq \min_{h' \in V} UB(h', Q, \delta/n)\}$
8.     $h_t \leftarrow \arg\min_{h \in V} UB(h, Q, \delta/n)$
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 $\hat{t} = \arg\min_{t \in \{1,2,\ldots,n\}} \beta_t$
```

Algorithm 1 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. As long as these bounds satisfy

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

for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ and any $\delta' \in (0,1)$, and UB and LB converge to each other as m grows, this algorithm is known to be correct, in that $er(\hat{h}) - \nu$ converges to 0 in probability [BBL06]. For instance, [BBL06] suggest defining these functions based on classic results on uniform convergence rates in passive learning [Vap82], such as

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

where $G(m, \delta') = \frac{1}{m} + \sqrt{\frac{\ln \frac{4}{\delta'} + d \ln \frac{2em}{d}}{m}}$, and by convention $G(0, \delta') = \infty$. This choice is justified by the following lemma, due to [Vap98].

Lemma 2 *For any distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$, 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*

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

◇

To avoid computational issues, instead of explicitly representing the sets V and R, we may implicitly represent them by a set of constraints imposed by the condition 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* examples.

The second algorithm we study was originally proposed by [DHM07]. It uses a type of constrained passive learning subroutine, LEARN, defined as follows.

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

If no $h \in \mathbb{C}$ has $er_{\mathcal{L}}(h) = 0$, define $\text{LEARN}_{\mathbb{C}}(\mathcal{L}, Q) = \varnothing$. Algorithm 2 is defined in terms of a function $\Delta_m(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta)$,

Algorithm 2
Input: hypothesis class $\mathbb{C}$, label budget n, confidence δ
Output: classifier $\hat{h}$, set of labeled examples $\mathcal{L}$, set of labeled examples Q

```
0. $\mathcal{L} \leftarrow \emptyset$, $Q \leftarrow \emptyset$
1. For $m = 1, 2, \ldots$
2.     If $|Q| = n$ or $|\mathcal{L}| = 2^n$, Return $\hat{h} = \textsc{Learn}_{\mathbb{C}}(\mathcal{L}, Q)$ along with $\mathcal{L}$ and $Q$
3.     For each $y \in \{-1, +1\}$, let $h^{(y)} = \textsc{Learn}_{\mathbb{C}}(\mathcal{L} \cup \{(X_m, y)\}, Q)$
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 $\mathcal{L} \leftarrow \mathcal{L} \cup \{(X_m, y)\}$
6.     Else Request the label $Y_m$ and let $Q \leftarrow Q \cup \{(X_m, Y_m)\}$
```

representing a threshold for a type of hypothesis test. This threshold must be set carefully, since the set $\mathcal{L} \cup Q$ is not actually an i.i.d. sample from $\mathcal{D}_{XY}$. [DHM07] suggest defining this function as

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

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., [DGL96]); this suggestion is based on a confidence bound they derive, and they prove the correctness of the algorithm with this definition. 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 fallback guarantees stating that neither algorithm is ever significantly worse than passive learning by empirical risk minimization [BBL06, DHM07]. 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.[2]

4.1 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 examples. In particular, [Han07] developed a technique for analyzing this type of algorithm, resulting in the following convergence rate guarantee for Algorithm 1.

Theorem 3 *[Han07] 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 $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

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

Similarly, the key to improvements from Algorithm 2 is that as m increases, we only need to request the labels of those examples 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 decreases, 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* examples 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. Using the disagreement coefficient, [DHM07] prove the following convergence rate guarantee.

Theorem 4 *[DHM07] 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 $\geq 1 - \delta$, $\forall n \in \mathbb{N}$, $er(\hat{h}_n) - \nu \leq$*

$$c\sqrt{\frac{\nu^2\theta d\log\frac{1}{\delta}\log\frac{n}{\theta\nu\delta}}{n}} + \sqrt{d\log\frac{1}{\delta}} \cdot exp\left\{-\sqrt{\frac{n}{c\theta d\log^2\frac{1}{\delta}}}\right\}. \qquad \diamond$$

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

[2] 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.

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.

4.2 Adaptation to 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 $\propto \nu n^{-1/2}$ [KÖ6]; 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 parameterization 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 version space decreases; thus, for small θ values, the region of disagreement should also be decreasing, 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, [CN06] 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 [CN06, Tsy04]. Furthermore, they 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. Later, in a personal communication, Langford and Castro claimed that Algorithm 1 also achieves this near-optimal rate (up to log factors) for the same learning problem (one-dimensional threshold classifiers under a uniform marginal distribution), leading to speculation that perhaps these improvements are achievable in the general case as well (under conditions on the disagreement coefficient).

Other than the one-dimensional threshold learning problem, it was not previously known whether Algorithm 1 or Algorithm 2 generally achieve convergence rates that exhibit these types of improvements.

4.3 Adaptive Rates in Active Learning

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 5 *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 $\geq 1 - \delta$,*

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

$\diamond$

Proof: 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.

First note that, by Lemma 2 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 the respective R set for that iteration, 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, $LB(h, Q, \delta/n) \le er(h|R) \le UB(h, Q, \delta/n)$, for the set R that Q is sampled under. Thus, we always have the invariant $\forall \gamma > 0, \{h \in V : er(h) - \nu \le \gamma\} \neq \emptyset$, and therefore also that $\forall t, er(h_t) - \nu = (er(h_t|R) - \inf_{h \in V} er(h|R))\mathbb{P}(R) \le \beta_t$. We will spend the remainder of the proof bounding the size of n sufficient to guarantee some $\beta_t \le \epsilon$.

Recalling the definition of the $h^{[k]}$ sequence (from Definition 1), note that after step 7,

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

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

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

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

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

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

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

By definition, every $h \in V$ has $LB(h, Q, \delta/n) \le \min_{h' \in V} UB(h', Q, \delta/n)$, so for this last set to be nonempty after step 7, we must have $\mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q|, \delta/n)$. On the other hand, if $\{h \in V : \limsup_k \mathbb{P}(h(X) \neq h^{[k]}(X)) > \mathbb{P}(R)/(2\theta)\} = \emptyset$, then $\mathbb{P}(DIS(V)) \le \mathbb{P}(DIS(\{h \in \mathbb{C} : \limsup_k \mathbb{P}(h(X) \neq h^{[k]}(X)) \le \mathbb{P}(R)/(2\theta)\}))$

$\leq \liminf_k \mathbb{P}(DIS(\{h \in \mathbb{C} : \mathbb{P}(h(X) \neq h^{[k]}(X)) \leq \mathbb{P}(R)/(2\theta)\})) \leq \liminf_k \theta_{h^{[k]}} \frac{\mathbb{P}(R)}{2\theta} = \frac{\mathbb{P}(R)}{2}$, so that we will definitely satisfy the condition in step 2 on the next round. Since $|Q|$ gets reset to 0 upon reaching step 3, we have that after every execution of step 7, $\mathbb{P}(R)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q|-1, \delta/n)$.

If $\mathbb{P}(R) \leq \frac{\epsilon}{2G(|Q|-1,\delta/n)} \leq \frac{\epsilon}{2G(|Q|,\delta/n)}$, then certainly $\beta_t \leq \epsilon$. So on any round for which $\beta_t > \epsilon$, we must have $\mathbb{P}(R) > \frac{\epsilon}{2G(|Q|-1,\delta/n)}$. Combined with the above observations, on any round with $\beta_t > \epsilon$, $\left(\frac{\epsilon}{2G(|Q|-1,\delta/n)}\right)^{\kappa-1}(2\mu\theta)^{-\kappa} < 4G(|Q|-1, \delta/n)$, which implies (by simple algebra)

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

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

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

suffices to guarantee some $\beta_t \leq \epsilon$. This implies the stated result by basic inequalities to bound the smallest value of ϵ satisfying (4) for a given value of n. ■

If the disagreement coefficient is small, Theorem 5 can represent a significant improvement in convergence rate compared to passive learning, where we typically expect rates of order $n^{-\kappa/(2\kappa-1)}$ [MT99, Tsy04, CN06]; this gap is especially notable when κ is small. In particular, the bound matches (up to log factors) the form of the minimax rate *lower bound* proven by [CN06] for threshold classifiers (where $\theta = 2$). Note that, unlike the analysis of [CN06], 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.

Note that, Theorem 5 is somewhat surprising, since the bounds UB and LB used to define the set V and the bounds β_t are not themselves adaptive to the noise conditions. Also note that, as before, n gets divided by θ^2 in the rates achieved by Algorithm 1. 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, hoping that as before, it is able to reduce the exponent of θ. 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.

$$\Delta_m(\mathcal{L}, Q, h^{(y)}, h^{(-y)}, \delta) = 3\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L} \cup Q, \delta; \mathcal{L}), \tag{5}$$

where $\hat{\mathcal{E}}_{\mathbb{C}}(\cdot, \cdot; \cdot)$ is defined in Appendix A, based on a notion of local Rademacher complexity studied by [Kol06]. Unlike the previous definitions, these definitions are known to be adaptive to Tsybakov's noise conditions, so that we would expect them to be asymptotically tighter and therefore allow the algorithm to more aggressively prune the set of candidate hypotheses. Using these definitions, we have the following theorem; its proof is included in Appendix B.

Theorem 6 *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 2 with threshold as in* (5), *when allowed n label requests and given 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, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu \leq \begin{cases} \frac{1}{\delta} \cdot exp\left\{-\sqrt{\frac{n}{cd\theta \log^3(d/\delta)}}\right\}, & \textit{when } \kappa = 1 \\ c\left(\frac{d\theta \log^2(dn/\delta)}{n}\right)^{\frac{\kappa}{2\kappa-2}}, & \textit{when } \kappa > 1 \end{cases}.$$

◇

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, but it is likely that this can be removed with a refined definition of $\hat{\mathcal{E}}_{\mathbb{C}}$, and therefore is not of fundamental significance. The bound in Theorem 6 is stated in terms of the VC dimension d. However, for certain nonparametric function classes (e.g., with $d = \infty$), it is sometimes preferable to quantify the complexity of the class in terms of a constraint on the *entropy* (with bracketing) of the class (see e.g., [vdVW96, Tsy04, Kol06, CN07]). 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 $\epsilon \in [0, 1]$, $\omega_{\mathbb{C}}(m, \epsilon) \leq \alpha \epsilon^{\frac{1-\rho}{2}} m^{-1/2}$.* ◇

In particular, as noted by [Kol06], the entropy with bracketing condition used in the original minimax analysis of [Tsy04] implies Condition 2. 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 [Kol06], and that this is sometimes tight [Tsy04]. The following theorem gives a bound on the rate of convergence of the same version of Algorithm 2 as in Theorem 6, this time in terms of the entropy with bracketing condition which, as before, is faster than the passive learning rate when the disagreement coefficient is small. The proof of this is included in Appendix B.

Theorem 7 *Suppose $\hat{h}_n$ is the classifier returned by Algorithm 2 with threshold as in* (5), *when allowed n label requests and given confidence parameter $\delta \in (0, 1/2)$. Suppose further that $\mathcal{D}_{XY}$ satisfies Conditions 1 and 2. Then there exists a finite (κ, μ, α and ρ -dependent) constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

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

◇

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 may be possible for Algorithm 1 as well.

5 Model Selection

While the previous sections address adaptation to the noise distribution, they are still restrictive in that they deal only with finite complexity hypothesis classes, where it is often unrealistic to expect convergence to the Bayes error rate to be achievable. 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 [Vap82]. 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 $\mathbb{C}_i$ can be defined so that it is always satisfied, under mild conditions on $\mathcal{X}$ (see e.g., [DGL96]). 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 constants $\mu_i > 0$ and $\kappa_i \geq 1$, such that $\forall \epsilon > 0, diam(\epsilon; \mathbb{C}_i) \leq \mu_i \epsilon^{\frac{1}{\kappa_i}}$.* $\diamond$

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., [Tsy04, Kol06]). In particular, [Kol06] 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, here referred to as Algorithm 3, is presented below, based on using Algorithm 2 as a subroutine. (It should also be possible to define an analogous method using Algorithm 1 as a subroutine instead.) The function $\hat{\mathcal{E}}_\cdot(\cdot, \cdot; \cdot)$ referred to in Algorithm 3 is defined in Appendix A.

This method can be shown to correctly converge in probability to an error rate of ν_∞ at a rate never significantly worse than the original passive learning method of [Kol06], as desired. Additionally, we have the following guarantee on the rate of convergence under Condition 3. The proof of this result, and the others in this section, are similar in style to Koltchinskii's original proofs, though some care is needed due to the altered sampling distribution and the constraint set $\mathcal{L}_{jn}$. However, these issues are addressed nicely by the several lemmas we have generated from the proofs of the previous section (in Appendix B). The details of these proofs are included in Appendix B.2.

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 Condition 3. Then there exist finite (κ_i and μ_i -dependent) constants c_i such that, with probability $\geq 1 - \delta, \forall n \in \mathbb{N}, er(\hat{h}_n) - \nu_\infty \leq$*

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

$\diamond$

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 6 for applying Algorithm 2 directly, with hypothesis class $\mathbb{C}_i$.

As in the case of finite-complexity $\mathbb{C}$, we can also show a variant of this result when the complexities are quantified in terms of the entropy with bracketing. Specifically, consider the following condition and theorem. Again, this represents an improvement over known results for passive learning when the disagreement coefficient is small.

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

Theorem 9 *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 $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

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

$\diamond$

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 version.

Theorem 10 *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 there exists a constant $\mu > 0$ such that for all measurable $h : \mathcal{X} \to \{-1, 1\}$, $er(h) - \nu^* \geq \mu \mathbb{P}\{h(X) \neq h^*(X)\}$. Then there exists a finite (μ-dependent) constant c such that, with probability $\geq 1 - \delta$, $\forall n \in \mathbb{N}$,*

$$er(\hat{h}_n) - \nu^* \leq c \min_{i \in \mathbb{N}} (\nu_i - \nu^*) + exp\left\{ -\sqrt{\frac{n}{c d_i \theta_i \log^3 \frac{i d_i}{\delta}}} \right\}.$$

$\diamond$

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.

```
Algorithm 3
Input: nested sequence of classes $\{\mathbb{C}_i\}$, label budget $n$, confidence parameter $\delta$
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 in (5), allowing $\lfloor n/(2i^2)\rfloor$ label requests, and confidence $\delta/(2i^2)$
2.     Let $h_{in} \leftarrow \textsc{Learn}_{\mathbb{C}_i}(\cup_{j\geq i}\mathcal{L}_{jn}, Q_{in})$
3.     If $h_{in} \neq \varnothing$ and $\forall j$ s.t. $i < j \leq \lfloor\sqrt{n/2}\rfloor$,
           $er_{\mathcal{L}_{jn}\cup Q_{jn}}(h_{in}) - er_{\mathcal{L}_{jn}\cup Q_{jn}}(h_{jn}) \leq \frac{3}{2}\hat{\mathcal{E}}_{\mathbb{C}_j}(\mathcal{L}_{jn}\cup Q_{jn}, \delta/(2j^2); \mathcal{L}_{jn})$
4.         $\hat{h}_n \leftarrow h_{in}$
5. Return $\hat{h}_n$
```

6 Conclusions

Under Tsybakov's noise conditions, active learning can offer improved asymptotic convergence rates compared to passive learning when the disagreement coefficient is small. 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.

A Definition of $\hat{\mathcal{E}}$

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 sequence of labeled examples $Q = \{(X'_i, Y'_i)\}$ as

$$R(f; Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\xi_i f(X'_i).$$

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

For any two sequences of labeled examples $\mathcal{L} = \{(X'_i, Y'_i)\}$ and $Q = \{(X''_i, Y''_i)\}$, define $\mathbb{C}[\mathcal{L}] = \{h \in \mathbb{C} : er_{\mathcal{L}}(h) = 0\}$,

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

let $\hat{D}_{\mathbb{C}}(\epsilon; \mathcal{L}, Q) = \sup_{h_1, h_2 \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}, Q)} \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{1}[h_1(X''_i) \neq h_2(X''_i)]$, and define $\hat{\phi}_{\mathbb{C}}(\epsilon; \mathcal{L}, Q) = \frac{1}{2}\sup_{h_1, h_2 \in \hat{\mathbb{C}}(\epsilon; \mathcal{L}, Q)} R(h_1 - h_2; Q)$. Let $\delta \in (0, 1]$, $m \in \mathbb{N}$, and define $s_m(\delta) = \ln\frac{20m^2\log_2(3m)}{\delta}$.

Let $\mathbb{Z}_\epsilon = \{j \in \mathbb{Z} : 2^j \geq \epsilon\}$, and for any sequence of labeled examples $Q = \{(X'_i, Y'_i)\}$, define

$$Q_m = \{(X'_1, Y'_1), (X'_2, Y'_2), \ldots, (X'_m, Y'_m)\}.$$

We use the following notation of Koltchinskii [Kol06] with only minor modifications. For $\epsilon \in [0, 1]$, define

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

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

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. We also define $\hat{\mathcal{E}}_{\mathbb{C}}(\emptyset, \delta; \mathbb{C}, \mathcal{L}) = \infty$ by convention.

B Main Proofs

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 $\hat{h}^*_m = \arg\min_{h \in \mathbb{C}} er_m(h)$ be the empirical risk minimizer in $\mathbb{C}$ for the *true* labels of the first m examples.

For $\epsilon > 0$, define $\mathbb{C}(\epsilon) = \{h \in \mathbb{C} : er(h) - \nu \leq \epsilon\}$. For $m \in \mathbb{N}$, let

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

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

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

where, for our purposes, we can take $\tilde{K} = 8272$ and $\tilde{c} = 3$. We also define $\tilde{\mathcal{E}}_{\mathbb{C}}(0, \delta) = \infty$. The following lemma is crucial to all of the proofs that follow.

Lemma 11 *[Kol06] There is an event $E_{\mathbb{C},\delta}$ with $\mathbb{P}(E_{\mathbb{C},\delta}) \geq 1 - \delta/2$ such that, on event $E_{\mathbb{C},\delta}$, $\forall m \in \mathbb{N}, \forall h \in \mathbb{C}, \forall \tau \in (0, 1/m), \forall h' \in \mathbb{C}(\tau)$,*

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

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

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

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

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

This lemma essentially follows from details of the proof of Koltchinskii's Theorem 1, Lemma 2, and Theorem 3 [Kol06][3]. We do not provide a proof of Lemma 11 here. The reader is referred to Koltchinskii's paper for the details.

[3] Our $\min_{m \leq |Q|}$ modification to Koltchinskii's version of $\hat{\mathcal{E}}_{\mathbb{C}}(m, \delta)$ is not a problem, since $\phi_{\mathbb{C}}(m, \epsilon)$ and $\frac{s_m(\delta)}{m}$ are nonincreasing functions of m.

B.1 Proofs Relating to Section 4

For $\ell \in \mathbb{N} \cup \{0\}$, let $\mathcal{L}^{(\ell)}$ and $Q^{(\ell)}$ denote the sets $\mathcal{L}$ and Q, respectively, in step 4 of Algorithm 2, when $m - 1 = \ell$; if this never happens during execution, define $\mathcal{L}^{(\ell)} = \emptyset, Q^{(\ell)} = \mathcal{Z}_\ell$.

Lemma 12 *On event* $E_{\mathbb{C},\delta}$, $\forall \ell \in \mathbb{N} \cup \{0\}$,

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

$$\textit{and } \forall \epsilon \geq \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta), \;\; \hat{h}^*_\ell \in \hat{\mathbb{C}}_\ell(\epsilon; \mathcal{L}^{(\ell)}) \subseteq \hat{\mathbb{C}}_\ell(\epsilon; \emptyset). \qquad \diamond$$

Proof:[Lemma 12] Throughout this proof, we assume the event $E_{\mathbb{C},\delta}$ occurs. We proceed by induction on ℓ, with the base case of $\ell = 0$ (which clearly holds). Suppose the statements are true for all $\ell' < \ell$. The case $\mathcal{L}^{(\ell)} = \emptyset$ is trivial, so assume $\mathcal{L}^{(\ell)} \neq \emptyset$. For the inductive step, suppose $h \in \hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \emptyset)$. Then for all $\ell' < \ell$, we have $er_\ell(h) - er_\ell(\hat{h}^*_\ell) \leq \hat{\mathcal{E}}_{\mathbb{C}}(\ell', \delta)$. In particular, by Lemma 11, this implies $er(h) - \nu \leq \max\left\{2(er_\ell(h) - er_\ell(\hat{h}^*_\ell)), \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)\right\}$ $\leq 2\hat{\mathcal{E}}_{\mathbb{C}}(\ell', \delta)$, and thus for any $h' \in \mathbb{C}$, $er_{\ell'}(h) - er_{\ell'}(h') \leq er_{\ell'}(h) - er_{\ell'}(\hat{h}^*_{\ell'}) \leq \frac{3}{2} \max\left\{er(h) - \nu, \hat{\mathcal{E}}_{\mathbb{C}}(\ell', \delta)\right\}$ $\leq 3\hat{\mathcal{E}}_{\mathbb{C}}(\ell', \delta) = 3\hat{\mathcal{E}}_{\mathbb{C}}(Q^{(\ell')}, \delta; \mathcal{L}^{(\ell')})$. Thus, we must have $er_{\mathcal{L}^{(\ell)}}(h) = 0$, and therefore $h \in \hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \mathcal{L}^{(\ell)})$. Since this is the case for all such h, we must have that

$$\hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \mathcal{L}^{(\ell)}) \supseteq \hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \emptyset). \tag{6}$$

In particular, this implies that $\hat{U}_{\mathbb{C}}(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta), \delta; \mathcal{L}^{(\ell)}, Q^{(\ell)}) \geq \hat{U}_{\mathbb{C}}(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta), \delta; \emptyset, \mathcal{Z}_\ell) > \frac{1}{16}\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$, where the last inequality follows from the definition of $\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$, (which is a power of 2). Thus, we must have $\hat{\mathcal{E}}_{\mathbb{C}}(Q^{(\ell)} \cup \mathcal{L}^{(\ell)}, \delta; \mathcal{L}^{(\ell)}) \geq \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$.

The relation in (6) also implies that $\hat{h}^*_\ell \in \hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \mathcal{L}^{(\ell)})$, and therefore $\forall \epsilon \geq \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$, $\hat{\mathbb{C}}_\ell(\epsilon; \mathcal{L}^{(\ell)}) \subseteq \hat{\mathbb{C}}_\ell(\epsilon; \emptyset)$, which implies $\forall \epsilon \geq \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$, $\hat{U}_{\mathbb{C}}(\epsilon, \delta; \mathcal{L}^{(\ell)}, Q^{(\ell)}) \leq \hat{U}_{\mathbb{C}}(\epsilon, \delta; \emptyset, \mathcal{Z}_\ell)$. But this means $\hat{\mathcal{E}}_{\mathbb{C}}(Q^{(\ell)} \cup \mathcal{L}^{(\ell)}, \delta; \mathcal{L}^{(\ell)}) \leq \hat{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)$. Therefore, we must have equality. Thus, the lemma follows by the principle of induction. ■

Lemma 13 *Suppose for any* $n \in \mathbb{N}$, $\hat{h}_n$ *is the classifier returned by Algorithm 2 with threshold as in* (5), *when allowed* n *label requests and given confidence parameter* $\delta > 0$, *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* $H_{\mathbb{C},\delta} \cap E_{\mathbb{C},\delta}$, $\forall n \in \mathbb{N}$,

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

and

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

$\diamond$

Proof:[Lemma 13] Once again, assume event $E_{\mathbb{C},\delta}$ occurs. By Lemma 11, $\forall \tau > 0$,

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

Letting $\tau \to 0$, and noting that $er_{\mathcal{L}}(\hat{h}^*_{m_n}) = 0$ (Lemma 12) 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 11. 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 nonnegative integer $\ell < m$, let I_ℓ be the indicator for the event that Algorithm 2 requests the label $Y_{\ell+1}$ and let $N_m = \sum_{\ell=0}^{m-1} I_\ell$. Additionally, let I'_ℓ be independent Bernoulli random variables with

$$\mathbb{P}[I'_\ell = 1] = \mathbb{P}\left\{DIS(\mathbb{C}(2\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)))\right\}.$$

Let $N'_m = \sum_{\ell=0}^{m-1} I'_\ell$. We have that

$$\begin{aligned}
&\mathbb{P}\left[\{I_\ell = 1\} \cap E_{\mathbb{C},\delta}\right] \\
&\leq \mathbb{P}\left[\{X_{\ell+1} \in DIS(\hat{\mathbb{C}}_\ell(\hat{\mathcal{E}}_{\mathbb{C}}(Q^{(\ell)} \cup \mathcal{L}^{(\ell)}, \delta; \mathcal{L}_i^{(\ell)}); \mathcal{L}^{(\ell)}))\} \cap E_{\mathbb{C},\delta}\right] \\
&\leq \mathbb{P}\left[\{X_{\ell+1} \in DIS(\hat{\mathbb{C}}_\ell(\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta); \emptyset))\} \cap E_{\mathbb{C},\delta}\right] \\
&\leq \mathbb{P}\left[DIS(\mathbb{C}(2\tilde{\mathcal{E}}_{\mathbb{C}}(\ell, \delta)))\right] = \mathbb{P}[I'_\ell = 1].
\end{aligned}$$

The second inequality is due to Lemmas 12 and 11, while the third inequality is due to Lemma 11. Note that

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

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 $E_{\mathbb{C},\delta}$) occuring with probability $\geq 1 - \delta$, for every $n \in \mathbb{N}$,

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

■

Lemma 14 *On event $H_{\mathbb{C},\delta} \cap E_{\mathbb{C},\delta}$ (of Lemmas 11 and 13), under Condition 1, $\forall n \in \mathbb{N}$,*

$$\tilde{\mathcal{E}}_{\mathbb{C}}(m_n, \delta) \leq \begin{cases} \frac{1}{\delta} \cdot exp\left\{-\sqrt{\frac{n}{cd\theta \log^3 \frac{d}{\delta}}}\right\}, & \textit{if } \kappa = 1 \\ c\left(\frac{d\theta \log^2(nd/\delta)}{n}\right)^{\frac{\kappa}{2\kappa-2}}, & \textit{if } \kappa > 1 \end{cases},$$

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

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

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

Proof:[**Lemma 14**] We begin with the first case (Condition 1 only).

We know that

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

for some constant K (see e.g., [MN06]). Noting that $\phi_{\mathbb{C}}(m, \epsilon) \leq \omega_{\mathbb{C}}(m, diam(\epsilon; \mathbb{C}))$, we have that

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

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

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

$$\tilde{\mathcal{E}}_{\mathbb{C}}(m, \delta) \leq K\left(\frac{d \log \frac{m}{\delta}}{m}\right)^{\frac{\kappa}{2\kappa-1}}. \tag{7}$$

Plugging this into the query bound, we have that

$$n \leq \log_2 \frac{4m_n^2}{\delta} + 2e\theta\left(2 + \int_1^{m_n-1} \mu(2K')^{\frac{1}{\kappa}}\left(\frac{d \log \frac{x}{\delta}}{x}\right)^{\frac{1}{2\kappa-1}}\right). \tag{8}$$

If $\kappa > 1$, (8) is at most $K''\theta m_n^{\frac{2\kappa-2}{2\kappa-1}} d \log \frac{m_n}{\delta}$, for some constant K'' (depending on κ and μ). This implies $m_n \geq K^{(3)}\left(\frac{n}{\theta d \log \frac{n}{\delta}}\right)^{\frac{2\kappa-1}{2\kappa-2}}$, for some constant $K^{(3)}$. Plugging this into (7) and using Lemma 13 completes the proof for this case.

On the other hand, if $\kappa = 1$, (8) is at most $K''\theta d \log^2 \frac{m_n}{\delta}$, for some constant K'' (depending on κ and μ). This implies $m_n \geq \delta exp\left\{K^{(3)}\sqrt{\frac{n}{\theta d}}\right\}$, for some constant $K^{(3)}$. Plugging this into (7), using Lemma 13, and simplifying the expression with a bit of algebra completes this case.

For the bound in terms of ρ, [Kol06] proves that $\tilde{\mathcal{E}}_{\mathbb{C}}(m, \delta) \leq$

$$K' \max\left\{m^{-\frac{\kappa}{2\kappa+\rho-1}}, \left(\frac{\log \frac{m}{\delta}}{m}\right)^{\frac{\kappa}{2\kappa-1}}\right\} \leq K'\left(\frac{\log \frac{m}{\delta}}{m}\right)^{\frac{\kappa}{2\kappa+\rho-1}}, \tag{9}$$

for some constant K' (depending on μ, α, and κ). Plugging this into the query bound, we have that

$$n \leq \log_2 \frac{4m_n^2}{\delta} + 2e\theta\left(2 + \int_1^{m_n-1} \mu(2K')^{\frac{1}{\kappa}}\left(\frac{\log \frac{x}{\delta}}{x}\right)^{\frac{1}{2\kappa+\rho-1}}\right)$$

$\leq K''\theta m_n^{\frac{2\kappa+\rho-2}{2\kappa+\rho-1}} \log \frac{m_n}{\delta}$, for some constant K'' (depending on κ, μ, α, and ρ). This implies $m_n \geq K^{(3)}\left(\frac{n}{\theta \log \frac{n}{\delta}}\right)^{\frac{2\kappa+\rho-1}{2\kappa+\rho-2}}$, for some constant $K^{(3)}$. Plugging this into (9) and using Lemma 13 completes the proof of this case. ■

Proof:[**Theorem 6** and **Theorem 7**] These theorems now follow directly from Lemmas 13 and 14. ■

B.2 Proofs Relating to Section 5

To simplify the notation in this section, define $\mathcal{L}Q_{in} = \mathcal{L}_{in} \cup Q_{in}$ for any $i \in \mathbb{N}, n \in \mathbb{N}$.

Lemma 15 *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

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

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:[**Lemma 15**] Continuing the notation from the proof of Lemma 12, for $\ell \in \mathbb{N} \cup \{0\}$, let $\mathcal{L}_{in}^{(\ell)}$ and $Q_{in}^{(\ell)}$ denote the sets $\mathcal{L}$ and Q, respectively, in step 4 of Algorithm 2, when $m - 1 = \ell$, when run with class $\mathbb{C}_i$, label budget $\lfloor n/(2i^2) \rfloor$, confidence parameter δ_i, and threshold as in (5); if $m - 1$ is never ℓ during execution, then define $\mathcal{L}_{in}^{(\ell)} = \emptyset$ and $Q_{in}^{(\ell)} = \mathcal{Z}_\ell$.

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 $h' \in \mathbb{C}_{i_n^*-1}(\tau_n) \cap \{h \in \mathbb{C}_i : er_{\mathcal{L}_{in}^{(\ell)}}(h) = 0\}$ but

$$er_\ell(h') - \min_{h \in \mathbb{C}_i} er_\ell(h) \geq er_\ell(h') - \min_{h \in \mathbb{C}_i : er_{\mathcal{L}_{in}^{(\ell)}}(h) = 0} er_\ell(h)$$

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

where the last equality is due to Lemma 12. Lemma 11 implies this will not happen for $i = i_n^* - 1$, so we can assume $i \geq i_n^*$. We therefore have (by Lemma 11) that

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

In particular, this implies that

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

Therefore,

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

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 $\forall n \in \mathbb{N}, h_{j_n^*n} \neq \emptyset$.

Now pick an arbitrary $i \in \mathbb{N}$ with $i > j = j_n^*$, and let $h' \in \mathbb{C}_j(\tau_n)$. Then $er_{\mathcal{L}Q_{in}}(h_{jn}) - er_{\mathcal{L}Q_{in}}(h_{in}) = er_{m_{in}}(h_{jn}) - er_{m_{in}}(h_{in})$

$$\begin{aligned}
\leq & \quad er_{m_{in}}(h_{jn}) - \min_{h\in\mathbb{C}_i} er_{m_{in}}(h) \\
\leq & \quad \frac{3}{2}\max\left\{er(h_{jn}) - \nu_i, \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i)\right\} \quad \text{(Lemma 11)} \\
= & \quad \frac{3}{2}\max\left\{er(h_{jn}) - \nu_j + \nu_j - \nu_i, \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i)\right\} \\
\leq & \quad \frac{3}{2}\max\begin{cases} 2(er_{m_{jn}}(h_{jn}) - er_{m_{jn}}(h') + \tau_n) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn},\delta_j) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i) \end{cases} \\
= & \quad \frac{3}{2}\max\begin{cases} \hat{\mathcal{E}}_{\mathbb{C}_j}(m_{jn},\delta_j) + \nu_j - \nu_i \\ \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i) \end{cases} \quad \text{(since } j \geq i_n^*\text{)} \\
= & \quad \frac{3}{2}\hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i) \quad \text{(by definition of } j_t^*\text{)} \\
= & \quad \frac{3}{2}\hat{\mathcal{E}}_{\mathbb{C}}(\mathcal{L}_{in} \cup Q_{in},\delta_i;\mathcal{L}_{in}) \quad \text{(by Lemma 12).}
\end{aligned}$$

■

Lemma 16 *On the event* $\bigcap_{i=1}^{\infty} E_{\mathbb{C}_i,\delta_i}$, $\forall n \in \mathbb{N}$,

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

Proof:[Lemma 16] Let $h'_n \in \mathbb{C}_{j_n^*}(\tau_n)$ for $\tau_n \in (0, 2^{-n})$, $n \in \mathbb{N}$. $er(\hat{h}_n) = er(h_{\hat{i}_n n})$

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

The first inequality follows from Lemma 11. The second inequality is due to Lemma 15 (i.e., $j_n^* \geq i_n^*$). Letting $\tau_n \to 0$ in this last line, and using the definition of $\hat{i}_n$, we have that $er(\hat{h}_n) - \nu_\infty$ is at most

$$\begin{aligned}
& \nu_{j_n^*} - \nu_\infty + \\
& \quad \max\left\{2\left(\frac{3}{2}\hat{\mathcal{E}}_{\mathbb{C}_{j_n^*}}(\mathcal{L}Q_{j_n^*n},\delta_{j_n^*};\mathcal{L}_{j_n^*n})\right), \hat{\mathcal{E}}_{\mathbb{C}_{j_n^*}}(m_{j_n^*n},\delta_{j_n^*})\right\} \\
= & \quad \nu_{j_n^*} - \nu_\infty + 3\hat{\mathcal{E}}_{\mathbb{C}_{j_n^*}}(m_{j_n^*n},\delta_{j_n^*}) \quad \text{(Lemma 12)} \\
\leq & \quad 3\min_i\left(\nu_i - \nu_\infty + \hat{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i)\right) \quad \text{(by definition of } j_n^*\text{)} \\
\leq & \quad 3\min_i\left(\nu_i - \nu_\infty + \tilde{\mathcal{E}}_{\mathbb{C}_i}(m_{in},\delta_i)\right) \quad \text{(Lemma 11).}
\end{aligned}$$

■

We are now ready for the proof of Theorems 8 and 9.

Proof:[Theorem 8 and **Theorem 9]** These theorems now follow directly from Lemmas 16 and 14. That is, Lemma 16 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 14 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 the 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. ■

Definition 17 *Define* $\mathring{c} = \tilde{c} + 1$, $\mathring{D}(\epsilon) = \lim_{j\to\infty} diam(\epsilon;\mathbb{C}_j)$, $\mathring{U}_{\mathbb{C}_i}(m,\epsilon,\delta_i)$

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

and

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

◇

Lemma 18 *For any* $m, i \in \mathbb{N}$,

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

◇

Proof:[Lemma 18] For $\epsilon > \nu_i - \nu_\infty$, $\tilde{U}_{\mathbb{C}_i}(m,\epsilon,\delta_i)$

$$\begin{aligned}
&= \tilde{K}\left(\phi_{\mathbb{C}_i}(m,\tilde{c}\epsilon) + \sqrt{\tfrac{s_m(\delta_i)diam(\tilde{c}\epsilon;\mathbb{C}_i)}{m}} + \tfrac{s_m(\delta_i)}{m}\right) \\
&\leq \tilde{K}\left(\omega_{\mathbb{C}_i}(m,diam(\tilde{c}\epsilon;\mathbb{C}_i)) + \sqrt{\tfrac{s_m(\delta_i)diam(\tilde{c}\epsilon;\mathbb{C}_i)}{m}} + \tfrac{s_m(\delta_i)}{m}\right).
\end{aligned}$$

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

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

In particular, this implies that

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

■

Proof:[**Theorem 10**] By the same argument that lead to (7), we have that

$$\mathring{\mathcal{E}}_{\mathbb{C}_i}(m, \delta_i) \leq K_2 \frac{d_i \log \frac{mi}{\delta}}{m},$$

for some constant K_2 (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, Lemma 16 implies that $\forall i, n \in \mathbb{N}$,

$$\begin{aligned}
&er(\hat{h}_n) - \nu^* \\
&\leq \min\left\{1, 3 \min_{i \in \mathbb{N}} \left(2(\nu_i - \nu_\infty) + \mathring{\mathcal{E}}_{\mathbb{C}_i}(m_{in}, \delta_i)\right)\right\} \\
&\leq K_3 \min_{i \in \mathbb{N}} \left((\nu_i - \nu^*) + \min\left\{1, \frac{d_i \log \frac{m_{in} i}{\delta}}{m_{in}}\right\}\right),
\end{aligned}$$

for some constant K_3.

Now take $i \in \mathbb{N}$. The label request bound of Lemma 13, along with Lemma 18, implies that

$$\begin{aligned}
&\lfloor n/(2i^2) \rfloor \leq \\
&\log \frac{8m_{in}^2 i^2}{\delta} + K_4 \theta_i \left(2 + \int_1^{m_{in}-1} \max\left\{\nu_i - \nu^*, \frac{d_i \log \frac{xi}{\delta}}{x}\right\} dx\right) \\
&\qquad \leq K_5 \theta_i \max\left\{(\nu_i - \nu^*) m_{in}, d_i \log^2(m_{in}) \log \frac{i}{\delta}\right\}
\end{aligned}$$

Let $\gamma_i(n) = \sqrt{\frac{n}{i^2 \theta_i d_i \log \frac{i}{\delta}}}$. Then

$$\begin{aligned}
&\frac{d_i \log \frac{m_{in} i}{\delta}}{m_{in}} \\
&\leq K_6 \left((\nu_i - \nu^*) \tfrac{1 + \gamma_i(n)}{\gamma_i(n)^2} + d_i \log \tfrac{i}{\delta} \left(1 + \gamma_i(n)\right) e^{-c_2 \gamma_i(n)}\right).
\end{aligned}$$

Thus, $\min\left\{1, \frac{d_i \log \frac{m_{in} i}{\delta}}{m_{in}}\right\}$

$$\leq \min\left\{1, K_7 \left((\nu_i - \nu^*) + d_i \log \tfrac{i}{\delta} \left(1 + \gamma_i(t)\right) e^{-c_2 \gamma_i(n)}\right)\right\}.$$

The result follows from this by some simple algebra. ■

Acknowledgments

I extend my sincere thanks to Larry Wasserman for numerous helpful discussions, and also to John Langford for initially pointing out to me the possibility of A^2 adapting to Tsybakov's noise conditions for threshold classifiers. I am also grateful for funding from the NSF Grant IIS-0713379 awarded to Eric Xing, and an IBM PhD Fellowship.

References

[BBL06] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proc. of the 23rd International Conference on Machine Learning*, 2006.

[CAL94] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[CN06] R.M. Castro and R.D. Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

[CN07] R.M. Castro and R.D. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

[DHM07] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

[Han07] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[KÖ6] M. Kääriäinen. Active learning in the non-realizable case. In *Proc. of the 17th International Conference on Algorithmic Learning Theory*, 2006.

[Kol06] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[MN06] P. Massart and E. Nedelec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[MT99] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

[Tsy04] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Vap82] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[Vap98] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.