A Bound on the Label Complexity of Agnostic Active Learning

Steve Hanneke SHANNEKE@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

We study the label complexity of pool-based active learning in the agnostic PAC model. Specifically, we derive general bounds on the number of label requests made by the A^2 algorithm proposed by Balcan, Beygelzimer & Langford (Balcan et al., 2006). This represents the first nontrivial general-purpose upper bound on label complexity in the agnostic PAC model.

1. Introduction

In active learning, a learning algorithm is given access to a large pool of unlabeled examples, and is allowed to request the label of any particular example from that pool. The objective is to learn an accurate classifier while requesting as few labels as possible. This contrasts with passive (semi)supervised learning, where the examples to be labeled are chosen randomly. In comparison, active learning can often significantly decrease the work load of human annotators by more carefully selecting which examples from the unlabeled pool should be labeled. This is of particular interest for learning tasks where unlabeled examples are available in abundance, but labeled examples require significant effort to obtain.

In the passive learning literature, there are well-known bounds on the number of training examples necessary and sufficient to learn a near-optimal classifier with high probability (i.e., the sample complexity) (Vapnik, 1998; Blumer et al., 1989; Kulkarni, 1989; Benedek & Itai, 1988; Long, 1995). This quantity depends largely on the VC dimension of the concept space being learned (in a distribution-independent analysis) or the metric entropy (in a distribution-dependent analysis). However, significantly less is presently known about the analogous quantity for active learning: namely, the *label complexity*, or number of label requests that are necessary and sufficient to learn. This knowledge gap is especially marked in the *agnostic* learning setting, where class labels can be noisy, and we have no assumption about the amount or type of noise. Building a thorough understanding of label complexity, along with the quantities on which it depends, seems essential to fully exploit the potential of active learning.

In the present paper, we study the label complexity by way of bounding the number of label requests made by a recently proposed active learning algorithm, A^2 (Balcan et al., 2006), which provably learns in the agnostic PAC model. The bound we find for this algorithm depends critically on a particular quantity, which we call the *disagreement coefficient*, depending on the concept space and example distribution. This quantity is often simple to calculate or bound for many concept spaces. Although we find that the bound we derive is not always tight for the label complexity, it represents a significant step forward, since it is the first nontrivial *general-purpose* bound on label complexity in the agnostic PAC model.

The rest of the paper is organized as follows. In Section 2, we briefly review some of the related literature, to place the present work in context. In Section 3, we continue with the introduction of definitions and notation. Section 4 discusses a variety of simple examples to help build intuition. Moving on in Section 5, we state and prove the main result of this paper: an upper bound on the number of label requests made by A^2, based on the disagreement coefficient. Following this, in Section 6, we prove a lower bound for A^2 with the same basic dependence on disagreement coefficient. We conclude in Section 7 with some open problems.

2. Background

The recent literature on the label complexity of active learning has been bringing us steadily closer to understanding the nature of this problem. Within that literature, there is a mix of positive and negative results, as well as a wealth of open problems.

Appearing in *Proceedings of the 24^th International Conference on Machine Learning*, Corvallis, OR, 2007. Revised 04/2011.

While studying the noise-free (realizable) setting, Dasgupta defines a quantity ρ called the *splitting index* (Dasgupta, 2005). ρ is dependent on the concept space, data distribution, and a (new) parameter τ he defines, as well as the target function itself. It essentially quantifies how easy it is to reduce the diameter of the concept space. He finds that under the assumption that there is no noise, roughly $\tilde{O}(\frac{d}{\rho})$ label requests are sufficient (where d is VC dimension), and $\Omega(\frac{1}{\rho})$ are necessary for learning (for respectively appropriate τ values). Thus, it appears that something like splitting index may be an important quantity to consider when bounding the label complexity. However, at present the only published analysis using splitting index is restricted to the noise-free (realizable) case. Additionally, one can construct simple examples where the splitting index is $O(1)$ (for $\tau = O(\epsilon^2)$), but agnostic learning requires $\Omega\left(\frac{1}{\epsilon}\right)$ label requests (even when the noise rate is zero). See Appendix A for an example of this. Thus, agnostic active learning seems to be a fundamentally more difficult problem than realizable active learning.

In studying the possibility of active learning in the presence of arbitrary classification noise, Balcan, Beygelzimer, & Langford propose the A^2 algorithm (Balcan et al., 2006). The strategy behind A^2 is to induce confidence intervals for the error rates of all concepts, and remove any concepts whose estimated error rate is larger than the smallest estimate to a statistically significant extent. This guarantees that with high probability we do not remove the best classifier in the concept space. The key observation that sometimes leads to improvements over passive learning is that, since we are only interested in *comparing* the error estimates, we do not need to request the label of any example whose label is not in dispute among the remaining classifiers. Balcan et al. analyze the number of label requests A^2 makes for some example concept spaces and distributions (notably linear separators under the uniform distribution on the unit sphere). However, other than fallback guarantees, they do not derive a general bound on the number of label requests, applicable to *any* concept space and distribution. This is the focus of the present paper.

In addition to the above results, there are a number of known lower bounds, than which there cannot be a learning algorithm guarateeing a number of label requests smaller. In particular, Kulkarni proves that, even if we allow *arbitrary* binary-valued queries and there is no noise, any algorithm that learns to accuracy $1-\epsilon$ can guarantee no better than $\Omega(\log N(2\epsilon))$ queries (Kulkarni et al., 1993), where $N(2\epsilon)$ is the size of a minimal 2ϵ-cover (defined below). Another known lower bound is due to Kääriäinen, who proves that in agnostic active learning, for most nontrivial concept spaces and distributions, if the noise rate is ν, then any algorithm that with probability $1-\delta$ outputs a classifier with error at most $\nu+\epsilon$ can guarantee no better than $\Omega\left(\frac{\nu^2}{\epsilon^2}\log\frac{1}{\delta}\right)$ label requests (Kääriäinen, 2006). In particular, these lower bounds imply that we can reasonably expect even the tightest general upper bounds on the label complexity to have some term related to $\log N(\epsilon)$ and some term related to $\frac{\nu^2}{\epsilon^2}\log\frac{1}{\delta}$.

3. Notation and Definitions

Let $\mathcal{X}$ be an *instance space*, comprising all possible examples we may ever encounter. $\mathbb{C}$ is a set of measurable functions $h : \mathcal{X} \to \{-1, 1\}$, known as the *concept space*. $\mathcal{D}_{XY}$ is any probability distribution on $\mathcal{X} \times \{-1, 1\}$. In the active learning setting, we draw $(X, Y) \sim \mathcal{D}_{XY}$, but the Y value is hidden from the learning algorithm until requested. For convenience, we will abuse notation by saying $X \sim \mathcal{D}$, where $\mathcal{D}$ is the marginal distribution of $\mathcal{D}_{XY}$ over $\mathcal{X}$; we then say the learning algorithm (optionally) *requests* the label Y of X (which was implicitly sampled at the same time as X); we may sometimes denote this label Y by $Oracle(X)$. For any $h \in \mathbb{C}$ and distribution $\mathcal{D}'$ over $\mathcal{X} \times \{-1, 1\}$, let $er_{\mathcal{D}'}(h) = \mathcal{P}r_{(X,Y)\sim\mathcal{D}'}\{h(X) \neq Y\}$, and for $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \{-1, 1\})^m$, $er_S(h) = \frac{1}{m}\sum_{i=1}^{m} |h(x_i) - y_i|/2$. When $\mathcal{D}' = \mathcal{D}_{XY}$ (the distribution we are learning with respect to), we abbreviate this by $er(h) = er_{\mathcal{D}_{XY}}(h)$. The *noise rate*, denoted ν, is defined as $\nu = \inf_{h\in\mathbb{C}} er(h)$. Our objective in agnostic active learning is to, with probability $\geq 1-\delta$, output a classifier h with $er(h) \leq \nu+\epsilon$ without making many label requests.

Let $\rho_{\mathcal{D}}(\cdot, \cdot)$ be the pseudo-metric on $\mathbb{C}$ induced by $\mathcal{D}$, s.t. $\forall h, h' \in \mathbb{C}, \rho_{\mathcal{D}}(h, h') = \mathcal{P}r_{X\sim\mathcal{D}}\{h(X) \neq h'(X)\}$. An ϵ-cover of $\mathbb{C}$ with respect to $\mathcal{D}$ is any set $V \subseteq \mathbb{C}$ such that $\forall h \in \mathbb{C}, \exists h' \in V : \rho_{\mathcal{D}}(h, h') \leq \epsilon$. We additionally let $N(\epsilon)$ denote the size of a minimal ϵ-cover of $\mathbb{C}$ with respect to $\mathcal{D}$. It is known that $N(\epsilon) < 2\left(\frac{2e}{\epsilon}\ln\frac{2e}{\epsilon}\right)^d$, where d is the VC dimension of $\mathbb{C}$ (Haussler, 1992). To focus on learnable cases, we assume $d < \infty$.

Definition 1. *For a set $V \subseteq \mathbb{C}$, define the* region of disagreement

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

Definition 2. *The* disagreement rate $\Delta(V)$ *of a set $V \subseteq \mathbb{C}$ is defined as*

$$\Delta(V) = \mathcal{P}r_{X\sim\mathcal{D}}\{X \in DIS(V)\}.$$

Definition 3. *For $h \in \mathbb{C}$, $r > 0$, let*

$$B(h,r) = \{h' \in \mathbb{C} : \rho_{\mathcal{D}}(h',h) \leq r\}$$

and define the disagreement rate at radius r

$$\Delta_r = \sup_{h \in \mathbb{C}} \Delta(B(h,r)).$$

Definition 4. *The* disagreement coefficient *is the infimum value of $\theta > 0$ such that $\forall r > \nu + \epsilon$,*

$$\Delta_r \leq \theta r.$$

The disagreement coefficient plays a critical role in the bounds of the following sections, which are increasing in this θ. Roughly speaking, it quantifies how quickly the region of disagreement can grow as a function of the radius of the version space.

4. Examples

The canonical example of the potential improvements in label complexity of active over passive learning is the *thresholds* concept space. Specifically, consider the concept space of thresholds t_z on the interval $[0,1]$ (for $z \in [0,1]$), such that $t_z(x) = +1$ iff $x \geq z$. Furthermore, suppose $\mathcal{D}$ is uniform on $[0,1]$. In this case, it is clear that the disagreement coefficient is at most 2, since the region of disagreement of $B(t_z, r)$ is roughly $\{x \in [0,1] : |x - z| \leq r\}$. That is, since the disagreement region grows at rate 1 in two disjoint directions as r increases, the disagreement coefficient $\theta = 2$.

As a second example, consider the disagreement coefficient for *intervals* on $[0,1]$. As before, let $\mathcal{X} = [0,1]$ and $\mathcal{D}$ be uniform, but this time $\mathbb{C}$ is the set of intervals $I_{[a,b]}$ such that for $x \in [0,1]$, $I_{[a,b]}(x) = +1$ iff $x \in [a,b]$ (for $a, b \in [0,1]$, $a \leq b$). In contrast to thresholds, the space of intervals serves as a canonical example of situations where active learning *does not help* compared to passive learning. This fact clearly shows itself in the disagreement coefficient, which is $\frac{1}{\nu+\epsilon}$ here, since $\Delta_r = 1$ for all $r > \nu + \epsilon$. To see this, note that the set $B(I_{[0,0]}, r)$ contains all concepts of the form $I_{[a,a]}$. Note that $\frac{1}{\nu+\epsilon}$ is the largest possible value for θ.

An interesting extension of the intervals example is the space of *p-intervals*, or all intervals $I_{[a,b]}$ such that $b - a \geq p \in ((\nu+\epsilon)/2, 1/8)$. These spaces span the range of difficulty, with active learning becoming easier as p increases. This is reflected in the θ value, since here $\theta = \frac{1}{2p}$. When $r < 2p$, every interval in $B(I_{[a,b]}, r)$ has its lower and upper boundaries within r of a and b, respectively; thus, $\Delta_r \leq 4r$. However, when $r \geq 2p$, *every* interval of width p is in $B(I_{[0,p]}, r)$, so $\Delta_r = 1$.

As an example that takes a (small) step closer to realistic learning scenarios, consider the following theorem.

Theorem 1. *If $\mathcal{X}$ is the surface of the origin-centered unit sphere in $\mathbb{R}^d$ for $d > 2$, $\mathbb{C}$ is the space of homogeneous linear separators*[1]*, and $\mathcal{D}$ is the uniform distribution on $\mathcal{X}$, then the disagreement coefficient θ satisfies*

$$\frac{1}{4}\min\left\{\pi\sqrt{d}, \frac{1}{\nu+\epsilon}\right\} \leq \theta \leq \min\left\{\pi\sqrt{d}, \frac{1}{\nu+\epsilon}\right\}.$$

Proof. First we represent the concepts in $\mathbb{C}$ as weight vectors $w \in \mathbb{R}^d$ in the usual way. For $w_1, w_2 \in \mathbb{C}$, by examining the projection of $\mathcal{D}$ onto the subspace spanned by $\{w_1, w_2\}$, we see that $\rho_{\mathcal{D}}(w_1, w_2) = \frac{\arccos(w_1 \cdot w_2)}{\pi}$. Thus, for any $w \in \mathbb{C}$ and $r \leq 1/2$, $B(w,r) = \{w' : w \cdot w' \geq \cos(\pi r)\}$. Some simple trigonometry gives us that

$$DIS(B(w,r)) = \{x \in \mathcal{X} : |x \cdot w| \leq \sin(\pi r)\}.$$

Let $A_d = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$ denote the surface area of the unit sphere in $\mathbb{R}^d$, and let $C_d(h) = \frac{1}{2}A_d I_{2h-h^2}\left(\frac{d-1}{2}, \frac{1}{2}\right)$ denote the surface area of a height-h spherical cap (Li, 2011), where $I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a+b)}{a\Gamma(a)\Gamma(b)}\int_0^{x^a}\left(1-u^{1/a}\right)^{b-1}\mathrm{d}u$ is the regularized incomplete beta function. Then we can express Δ_r as

$$\Delta_r = 1 - \frac{2C_d(1-\sin(\pi r))}{A_d} = 1 - I_{\cos^2(\pi r)}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

As $I_x(a,b) = 1 - I_{1-x}(b,a)$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, this equals

$$\frac{2\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}\int_0^{\sin(\pi r)}\left(1-x^2\right)^{\frac{d-3}{2}}\mathrm{d}x. \qquad (*)$$

As $\sqrt{\frac{d}{3}} \leq \frac{2\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)} \leq \sqrt{d-2}$, we see $(*)$ is at most $\sqrt{d-2}\sin(\pi r) \leq \sqrt{d}\pi r$.

For the lower bound, $\Delta_{1/2} = 1 \Rightarrow \theta \geq \min\left\{2, \frac{1}{\nu+\epsilon}\right\}$, so we need only consider $\nu + \epsilon < \frac{1}{8}$. Supposing $\nu + \epsilon < r < \frac{1}{8}$, we have that $(*)$ is at least

$$\sqrt{\frac{d}{3}}\int_0^{\sin(\pi r)}\left(1-x^2\right)^{\frac{d}{2}}\mathrm{d}x \geq \sqrt{\frac{\pi}{3}}\int_0^{\sin(\pi r)}\sqrt{\frac{d}{\pi}}e^{-d\cdot x^2}\mathrm{d}x$$
$$\geq \frac{1}{2}\min\left\{\frac{1}{2}, \sqrt{d}\sin(\pi r)\right\} \geq \frac{1}{4}\min\left\{1, \sqrt{d}\pi r\right\}. \qquad \square$$

Given knowledge of the disagreement coefficient for $\mathbb{C}$ under $\mathcal{D}$, the following lemma allows us to extend this to a bound for any $\mathcal{D}'$ λ-close to $\mathcal{D}$. The proof is straightforward, and left as an exercise.

[1]Homogeneous linear separators are those that pass through the origin.

Input: concept space $\mathbb{C}$, accuracy parameter $\epsilon \in (0,1)$, confidence parameter $\delta \in (0,1)$
Output: classifier $\hat{h} \in \mathbb{C}$
Let $\hat{n} = \log_2\left(\frac{64}{\epsilon^2}\left(d \ln \frac{8}{\epsilon} + \ln \frac{8}{\epsilon\delta}\right)\right) \log_2 \frac{4}{\epsilon}$, and let $\delta' = \delta/\hat{n}$
0. $V_0 \leftarrow \mathbb{C}$, $S_0 \leftarrow \varnothing$, $i \leftarrow 0$, $j_1 \leftarrow 0$, $k \leftarrow 1$
1. While $\Delta(V_i)\left(\min_{h \in V_i} UB(S_i, h, \delta') - \min_{h \in V_i} LB(S_i, h, \delta')\right) > \epsilon$
2. $\quad V_{i+1} \leftarrow \{h \in V_i : LB(S_i, h, \delta') \le \min_{h' \in V_i} UB(S_i, h', \delta')\}$
3. $\quad i \leftarrow i + 1$
4. $\quad$ If $\Delta(V_i) < \frac{1}{2}\Delta(V_{j_k})$
5. $\quad\quad k \leftarrow k + 1$; $j_k \leftarrow i$
6. $\quad S'_i \leftarrow$ Rejection sample $2^{i - j_k}$ samples x from $\mathcal{D}$ satisfying $x \in DIS(V_i)$
7. $\quad S_i \leftarrow \{(x, Oracle(x)) : x \in S'_i\}$
8. Return $\hat{h} = \arg\min_{h \in V_i} UB(S_i, h, \delta')$

Figure 1. The A^2 algorithm.

Lemma 1. *Suppose $\mathcal{D}'$ is such that, $\exists \lambda \in (0,1]$ s.t. for all measurable sets $A \subseteq \mathcal{X}$, $\lambda \mathcal{D}(A) \le \mathcal{D}'(A) \le \frac{1}{\lambda}\mathcal{D}(A)$. If Δ_r, θ, Δ'_r, and θ' are the disagreement rates at radius r and disagreement coefficients for $\mathcal{D}$ and $\mathcal{D}'$ respectively, then $\lambda \Delta_{\lambda r} \le \Delta'_r \le \frac{1}{\lambda}\Delta_{r/\lambda}$, and thus*

$$\lambda^2 \theta \le \theta' \le \frac{1}{\lambda^2}\theta.$$

5. Upper Bounds for the A^2 Algorithm

To prove bounds on the label complexity, we will additionally need to use some known results on finite sample rates of uniform convergence.

Definition 5. *Let d be the VC dimension of $\mathbb{C}$. For $m \in \mathbb{N}$, and $S \in (\mathcal{X} \times \{-1,1\})^m$, define*

$$G(m, \delta) = \frac{1}{m} + \sqrt{\frac{\ln \frac{4}{\delta} + d \ln \frac{2em}{d}}{m}}.$$

$$UB(S, h, \delta) = \min\{er_S(h) + G(|S|, \delta), 1\},$$
$$LB(S, h, \delta) = \max\{er_S(h) - G(|S|, \delta), 0\}.$$

By convention, $G(0, \delta) = 1$. The following lemma is due to Vapnik (Vapnik, 1998).

Lemma 2. *For any distribution $\mathcal{D}_i$ over $\mathcal{X} \times \{-1,1\}$, and any $m \in \mathbb{N}$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}_i^m$, every $h \in \mathbb{C}$ satisfies*

$$|er_S(h) - er_{\mathcal{D}_i}(h)| \le G(m, \delta).$$

In particular, this means

$$er_{\mathcal{D}_i}(h) - 2G(|S|, \delta) \le LB(S, h, \delta) \le$$
$$er_{\mathcal{D}_i}(h) \le UB(S, h, \delta) \le er_{\mathcal{D}_i}(h) + 2G(|S|, \delta).$$

Furthermore, for $\gamma > 0$, if $m \ge \frac{4}{\gamma^2}\left(2d \ln \frac{4}{\gamma} + \ln \frac{4}{\delta}\right)$, then $G(m, \delta) < \gamma$.

We use a (somewhat simplified) version of the A^2 algorithm, presented by Balcan et. al (Balcan et al., 2006). The algorithm is given in Figure 1.

The motivation behind the A^2 algorithm is to maintain a set of concepts V_i that we are confident contains any concepts with minimal error rate. If we can guarantee with statistical significance that a concept $h_1 \in V_i$ has error rate worse than another concept $h_2 \in V_i$, then we can safely remove the concept h_1 since it is suboptimal. To achieve such a statistical guarantee, the algorithm employs two-sided confidence intervals on the error rates of each classifier in the concept space; however, since we are only interested in the relative *differences* between error rates, on each iteration we obtain this confidence interval for the error rate when $\mathcal{D}$ is restricted to the *region of disagreement* $DIS(V_i)$. This restriction to the region of disagreement is the primary source of any improvements A^2 achieves over passive learning. We measure the progress of the algorithm by the reduction in the disagreement rate $\Delta(V_i)$; the key question in studying the number of label requests is bounding the number of random labeled examples from the region of disagreement that are sufficient to remove enough concepts from V_i to significantly reduce the measure of the region of disagreement.

Theorem 2. *If θ is the disagreement coefficient for $\mathbb{C}$, then with probability at least $1 - \delta$, given the inputs $\mathbb{C}$, ϵ, and δ, A^2 outputs $\hat{h} \in \mathbb{C}$ with $er(\hat{h}) \le \nu + \epsilon$, and the number of label requests made by A^2 is at most*

$$O\left(\theta^2\left(\frac{\nu^2}{\epsilon^2} + 1\right)\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right) \log \frac{1}{\epsilon}\right).$$

Proof. Let κ be the value of k and ι be the value of i when the algorithm halts. By convention, let $j_{\kappa+1} = \iota + 1$. Let $\gamma_i = \max_{h \in V_i}(UB(S_i, h, \delta') - LB(S_i, h, \delta'))$. Since having $\gamma_i \le \epsilon$ would break the loop at step 1, Lemma 2 implies we always have $|S_i| \leq \frac{16}{\epsilon^2}\left(2d\ln\frac{8}{\epsilon}+\ln\frac{4}{\delta'}\right)$, and thus $\iota \leq (\kappa+1)\log_2\left(\frac{16}{\epsilon^2}\left(2d\ln\frac{8}{\epsilon}+\ln\frac{4}{\delta'}\right)\right)$. $\Delta(V_i) \leq \epsilon$ also suffices to break from the loop, so $\kappa \leq \log_2\frac{2}{\epsilon}$. Thus, $\iota \leq \hat{n}$. Lemma 2 and a union bound imply that, with probability $\geq 1-\delta$, for every i and every $h \in \mathbb{C}$, $|er_{S_i}(h) - er_{\mathcal{D}_i}(h)| \leq G(|S_i|, \delta')$, where $\mathcal{D}_i$ is the conditional distribution of $\mathcal{D}_{XY}$ given that $X \in DIS(V_i)$. For the remainder of this proof, we assume that these inequalities hold for all such S_i and $h \in \mathbb{C}$. In particular, this means we never remove the best classifier from V_i. Additionally, $\forall h_1, h_2 \in V_i$ we must have $\Delta(V_i)(er_{\mathcal{D}_i}(h_1) - er_{\mathcal{D}_i}(h_2)) = er(h_1) - er(h_2)$. Combined with the nature of the halting criterion, this implies that $er(\hat{h}) \leq \nu + \epsilon$, as desired.

The rest of the proof bounds the number of label requests made by A^2. Let $h^* \in V_i$ be such that $er(h^*) \leq \nu+\epsilon$. We consider two cases: large and small $\Delta(V_i)$. Informally, when $\Delta(V_i)$ is relatively large, the concepts far from h^* are responsible for most of the disagreements, and since these must have relatively large error rates, we need only a few examples to remove them. On the other hand, when $\Delta(V_i)$ is small, the halting condition is easy to satisfy.

We begin with the case where $\Delta(V_i)$ is large. Specifically, let $i' = \max\{i \leq \iota : \Delta(V_i) > 8\theta(\nu+\epsilon)\}$. (If no such i' exists, we can skip this case). Then $\forall i \leq i'$, let

$$V_i^{(\theta)} = \left\{h \in V_i : \rho_{\mathcal{D}}(h, h^*) > \frac{\Delta(V_i)}{2\theta}\right\}.$$

Since for $h \in V_i$, $\rho_{\mathcal{D}}(h,h^*)/\Delta(V_i) \leq er_{\mathcal{D}_i}(h) + er_{\mathcal{D}_i}(h^*) \leq er_{\mathcal{D}_i}(h) + \frac{\nu+\epsilon}{\Delta(V_i)}$, we have

$$\begin{aligned} V_i^{(\theta)} &\subseteq \left\{h \in V_i : er_{\mathcal{D}_i}(h) > \frac{1}{2\theta} - \frac{\nu+\epsilon}{\Delta(V_i)}\right\} \\ &\subseteq \left\{h \in V_i : er_{\mathcal{D}_i}(h) - \frac{1}{8\theta} > er_{\mathcal{D}_i}(h^*) + \frac{3}{8\theta} - 2\frac{\nu+\epsilon}{\Delta(V_i)}\right\} \\ &\subseteq \left\{h \in V_i : er_{\mathcal{D}_i}(h) - \frac{1}{8\theta} > er_{\mathcal{D}_i}(h^*) + \frac{1}{8\theta}\right\}. \end{aligned}$$

Let $\bar{V}_i$ denote the latter set. By Lemma 2, S_i of size $O\left(\theta^2\left(d\log\theta + \log\frac{1}{\delta'}\right)\right)$ suffices to guarantee every $h \in \bar{V}_i$ has $LB(S_i, h, \delta') > UB(S_i, h^*, \delta')$ in step 2. $V_i^{(\theta)} \subseteq \bar{V}_i$ and $\Delta(V_i \setminus V_i^{(\theta)}) \leq \Delta_{\frac{\Delta(V_i)}{2\theta}} \leq \frac{1}{2}\Delta(V_i)$, so in particular, any value of k for which $j_k \leq i'+1$ satisfies $|S_{j_k-1}| = O\left(\theta^2\left(d\log\theta + \log\frac{1}{\delta'}\right)\right)$.

To handle the remaining case, suppose $\Delta(V_i) \leq 8\theta(\nu+\epsilon)$. In this case, S_i of size $O\left(\theta^2\frac{(\nu+\epsilon)^2}{\epsilon^2}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta'}\right)\right)$ suffices to make $\gamma_i \leq \frac{\epsilon}{\Delta(V_i)}$, satisfying the halting condition. Therefore, every k for which $j_k > i'+1$ satisfies $|S_{j_k-1}| = O\left(\theta^2\frac{(\nu+\epsilon)^2}{\epsilon^2}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta'}\right)\right)$.

Since for $k>1$, $\sum_{i=j_{(k-1)}}^{j_k-1}|S_i| \leq 2|S_{j_k-1}|$, we have that $\sum_{i=1}^{\iota}|S_i| = O\left(\theta^2\frac{(\nu+\epsilon)^2}{\epsilon^2}\left(d\log\frac{1}{\epsilon}+\log\frac{1}{\delta'}\right)\kappa\right)$. Noting that $\kappa = O(\log\frac{1}{\epsilon})$ and $\log\frac{1}{\delta'} = O\left(d\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)$ completes the proof. □

Note that we can get an easy improvement to the bound by replacing $\mathbb{C}$ with an $\frac{\epsilon}{2}$-cover of $\mathbb{C}$, using bounds for a finite concept space instead of VC bounds, and running the algorithm with accuracy parameter $\frac{\epsilon}{2}$. This yields a similar, but sometimes much tighter, label complexity bound of

$$O\left(\theta^2\left(\frac{\nu^2}{\epsilon^2}+1\right)\log\frac{N(\epsilon/2)\log\frac{1}{\epsilon}}{\delta}\log\frac{1}{\epsilon}\right).$$

6. Lower Bounds for the A^2 Algorithm

In this section, we prove a lower bound on the worst-case number of label requests made by A^2. As mentioned in Section 2, there are known lower bounds of $\Omega\left(\frac{\nu^2}{\epsilon^2}\log\frac{1}{\delta}\right)$ and $\Omega(\log N(2\epsilon))$, than which *no* algorithm can guarantee better (Kulkarni et al., 1993; Kääriäinen, 2006). However, this leaves open the question of whether the θ^2 factor in the bound is necessary. The following theorem shows that it is for A^2.

Theorem 3. *For any $\mathbb{C}$ and $\mathcal{D}$, there exists an oracle with $\nu = 0$ such that, if θ is the disagreement coefficient, with probability $1-\delta$, the version of A^2 presented above makes a number of label requests at least*

$$\Omega\left(\theta^2\left(d\log\theta + \log\frac{1}{\delta}\right)\right).$$

Proof. The bound clearly holds if $\theta = 0$, so assume $\theta > 0$. By definition of disagreement coefficient, there is some $\alpha_0 > 0$ such that $\forall\alpha \in (0,\alpha_0)$, $\exists r_\alpha \in (\epsilon, 1], h_\alpha \in \mathbb{C}$ such that $\Delta(B(h_\alpha, r_\alpha)) \geq \Delta_{r_\alpha} - \alpha \geq \theta r_\alpha - 2\alpha > 0$. For some such α, let $Oracle(x) = h_\alpha(x)$ for all $x \in \mathcal{X}$. Clearly $\nu = 0$. As before, we assume all bound evaluations in the algorithm are valid, which occurs with probability $\geq 1-\delta$. Since $LB(S_i, h_\alpha, \delta') = 0$ and $UB(S_i, h_\alpha, \delta') = G(|S_i|, \delta')$, if A^2 halts without removing any $h \in B(h_\alpha, r_\alpha)$, then $\exists i : UB(S_i, h_\alpha, \delta') \leq \frac{\epsilon}{\Delta(B(h_\alpha, r_\alpha))} \leq \frac{\epsilon}{\theta r_\alpha - 2\alpha} \leq \frac{r_\alpha}{\theta r_\alpha - 2\alpha}$. On the other hand, suppose A^2 removes some $h \in B(h_\alpha, r_\alpha)$ before halting, and in particular suppose the first time this happens is for some set S_i. In this case, $UB(S_i, h_\alpha, \delta') < LB(S_i, h, \delta') \leq er_{\mathcal{D}_i}(h) \leq \frac{er(h)}{\Delta(B(h_\alpha, r_\alpha))} \leq \frac{r_\alpha}{\theta r_\alpha - 2\alpha}$. In either case, by definition of $G(|S_i|, \delta')$, we must have $|S_i| = \Omega\left(\left(\theta - \frac{2\alpha}{r_\alpha}\right)^2\left(d\log\left(\theta - \frac{2\alpha}{r_\alpha}\right) + \log\frac{1}{\delta'}\right)\right)$. Since this is true for any such α, taking the limit as $\alpha \to 0$ proves the bound. □

Theorems 2 and 3 show that the variation in worst-case number of label requests made by A^2 for different $\mathbb{C}$ and $\mathcal{D}$ is largely determined by the disagreement coefficient (and VC dimension). Furthermore, they give us a good estimate of the number of label requests made by A^2. One natural question to ask is whether Theorem 2 is also tight for the *label complexity* of the learning problem. The following example indicates this is not the case. In particular, this means that A^2 can sometimes be suboptimal.

Suppose $\mathcal{X} = [0,1]^n$, and $\mathbb{C}$ is the space of axis-aligned rectangles on $\mathcal{X}$. That is, each $h \in \mathbb{C}$ can be expressed as n pairs $((a_1,b_1),(a_2,b_2),\ldots,(a_n,b_n))$, such that $\forall x \in \mathcal{X}, h(x) = 1$ iff $\forall i, a_i \leq x_i \leq b_i$. Furthermore, suppose $\mathcal{D}$ is the uniform distribution on $\mathcal{X}$. We see immediately that $\theta = \frac{1}{\epsilon+\nu}$, since $\forall r > 0, \Delta_r = 1$. We will show the bound is not tight for the case when $\nu = 0$.[2] In this case, the bound value is $\Omega\left(\frac{1}{\epsilon^2}\left(n \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$.

Theorem 4. *When $\nu = 0$, the agnostic active learning label complexity of axis-aligned rectangles on $[0,1]^n$ with respect to the uniform distribution is at most*

$$O\left(n \log \frac{n}{\epsilon\delta} + \frac{1}{\epsilon} \log \frac{1}{\delta}\right).$$

A proof sketch for Theorem 4 is included in Appendix B. This clearly shows that the bound based on A^2 is sometimes not tight with respect to the true label complexity of learning problems. Furthermore, when $\epsilon < \frac{1}{en}$, this problem has $\log N(\epsilon/2) \geq n$, so the improvements offered by learning with an $\frac{\epsilon}{2}$-cover cannot reduce the slack by much here (see Lemma 3 in Appendix B).

7. Open Problems

Whether or not one can modify A^2 in a general way to improve this bound is an interesting open problem. One possible strategy would be to use Occam bounds, and adaptively set the prior for each iteration, while also maintaining several different types of bounds simultaneously. However, it seems that in order to obtain the dramatic improvements needed to close the gap demonstrated by Theorem 4, we need a more aggressive strategy than sampling randomly from $DIS(V_i)$. For example, Balcan, Broder & Zhang (Balcan et al., 2007) present an algorithm for linear separators which samples from a carefully chosen subregion of $DIS(V_i)$. Though their analysis is for a restricted noise model, we might hope a similar idea is possible in the agnostic model. The end of Appendix A contains another interesting example that highlights this issue.

One important aspect of active learning that has not been addressed here is the value of unlabeled examples. Specifically, given an overabundance of unlabeled examples, can we use them to decrease the number of label requests required, and by how much? The splitting index bounds of Dasgupta (Dasgupta, 2005) can be used to study these types of questions in the noise-free setting; however, we have yet to see a thorough exploration of the topic for agnostic learning, where the role of unlabeled examples appears fundamentally different (at least in A^2).

[2]In this particular case, the agnostic label complexity with $\nu = 0$ is within constant factors of the realizable complexity. However, in general, agnostic learning with $\nu = 0$ is *not* the same as realizable learning, since we are still interested in algorithms that would tolerate noise if it were present. See Appendix A for an interesting example.

Acknowledgments

I am grateful to Nina Balcan for helpful discussions, and to Ye Nan for pointing out a mistake in the original proof of Theorem 1.

This research was sponsored through a generous grant from the Commonwealth of Pennsylvania. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring body, or other institution or entity.

References

Balcan, M.-F., Beygelzimer, A., & Langford, J. (2006). Agnostic active learning. *Proc. of the 23rd International Conference on Machine Learning.*

Balcan, M.-F., Broder, A., & Zhang, T. (2007). Margin based active learning. *Proc. of the 20th Conference on Learning Theory.*

Benedek, G., & Itai, A. (1988). Learnability by fixed distributions. *Proc. of the First Workshop on Computational Learning Theory* (pp. 80–90).

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, *36*, 929–965.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. *Advances in Neural Information Processing Systems 18.*

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, *100*, 78–150.

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

Kulkarni, S. R. (1989). *On metric entropy, vapnik-chervonenkis dimension, and learnability for a class of distributions* (Technical Report CICS-P-160). Center for Intelligent Control Systems.

Kulkarni, S. R., Mitter, S. K., & Tsitsiklis, J. N. (1993). Active learning using arbitrary binary valued queries. *Machine Learning*, *11*, 23–35.

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, *4*, 66–70.

Long, P. M. (1995). On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, *6*, 1556–1559.

Vapnik, V. (1998). *Statistical learning theory.* John Wiley & Sons, Inc.

A. Realizable vs. Agnostic with $\nu = 0$

The following example indicates that agnostic active learning with $\nu = 0$ is sometimes fundamentally more difficult than realizable learning.

Let $\epsilon < 1/4$, $N = \left\lfloor \frac{1}{2\epsilon} \right\rfloor$. Let $\mathcal{X} = \mathbb{Z}$, and define $\mathcal{D}$ such that, for $x \in \mathcal{X} : 0 < x \leq N$, $\mathcal{D}(x) = \frac{\epsilon}{4N}$ and $\mathcal{D}(-x) = \frac{1-\epsilon/4}{N}$. $\mathcal{D}$ has zero probability elsewhere. In particular, note that $\frac{3}{2}\epsilon < \mathcal{D}(-x) \leq 4\epsilon$ and $\frac{\epsilon^2}{2} \leq \mathcal{D}(x) \leq \epsilon^2$.

Define concept space $\mathbb{C} = \{h_1, h_2, \ldots\}$, where $\forall i, j \in \{1, 2, \ldots\}$, $h_i(0) = -1$ and

$$\begin{aligned} h_i(-j) &= 2I[i = j] - 1 \\ h_i(j) &= 2I[j \geq i] - 1. \end{aligned}$$

Note that this creates a learning problem where informative examples exist ($x \in \{1, \ldots, N\}$) but are rare.

Theorem 5. *For the learning problem described above, the realizable active learning label complexity is* $O\left(\log \frac{1}{\epsilon}\right)$.

Proof. By Chernoff and union bounds, drawing $\Theta\left(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon\delta}\right)$ unlabeled examples suffices to guarantee, with probability at least $1-\delta$, we have at least one unlabeled example of x, for all $x \in \{1, 2, \ldots, N\}$; suppose this happens. Suppose $f \in \mathbb{C}$ is the target function. If $f \notin \{h_1, h_2, \ldots, h_N\}$, querying the label of $x = N$ suffices to show $er(h_{N+1}) = 0$, so we output h_{N+1}. On the other hand, if we find $f(N) = +1$, we can perform binary search among $\{1, 2, \ldots, N\}$ to find the smallest $i > 0$ such that $f(i) = +1$. In this case, we must have $h_i = f$, so we output h_i after $O(\log N)$ queries. □

Theorem 6. *For the learning problem described above, any agnostic active learning algorithm requires* $\Omega\left(\frac{1}{\epsilon}\right)$ *label requests, even if the oracle always agrees with some* $f \in \mathbb{C}$, *(i.e., even if* $\nu = 0$*).*

Proof. Suppose A is a correct agnostic learning algorithm. The idea of the proof is to assume A is guaranteed to make fewer than $(1-2\delta)N$ queries with probability $\geq 1-\delta$ when the target function is some particular $f \in \mathbb{C}$, and then show that by adding noise we can force A to output a concept with error more than ϵ-worse than optimal with probability $> \delta$. Thus, either A cannot guarantee fewer than $(1-2\delta)N$ queries for that particular f, or A is not a correct agnostic learning algorithm.

Specifically, suppose that when the target function $f = h_{N+1}$, with probability $\geq 1-\delta$ A returns an ϵ-good concept after making $\leq q < (1-2\delta)N$ label requests. If A is successful, then whatever concept it outputs labels all of $\{-1, -2, \ldots, -N\}$ as -1. So in particular, letting the random variable $R = (R_1, R_2, \ldots)$ denote the sequence of examples A requests the labels of when *Oracle* agrees with h_{N+1}, this implies that with probability at least $1-\delta$, if $Oracle(R_i) = h_{N+1}(R_i)$ for $i \in \{1, 2, \ldots, \min\{q, |R|\}\}$, then A outputs a concept labeling all of $\{-1, -2, \ldots, -N\}$ as -1.

Now suppose instead of h_{N+1}, we pick the target function f' as follows. Let f' be identical to h_{N+1} on all of $\mathcal{X}$ except a single $x \in \{-1, -2, \ldots, -N\}$ where $f'(x) = +1$; the value of x for which this happens is chosen uniformly at random from $\{-1, -2, \ldots, -N\}$. Note that $f' \notin \mathbb{C}$. Also note that any concept in $\mathbb{C}$ other than h_{-x} is $> \epsilon$-worse than h_{-x}.

Now consider the behavior of A when Oracle answers queries with this f' instead of h_{N+1}. Let $Q = (Q_1, Q_2, \ldots)$ denote the random sequence of examples A queries the labels of when *Oracle* agrees with f'. In particular, note that if $R_i \neq x$ for $i \leq \min\{q, |R|\}$, then $Q_i = R_i$ for $i \leq \min\{q, |Q|\}$.

$$\begin{aligned} &\mathbb{E}_{f'}\left[\mathcal{P}r\{A \text{ outputs } h_{-x}\}\right] \\ &\qquad \leq \mathbb{E}_R\left[\mathcal{P}r_x\{\exists i \leq q : R_i = x\}\right] + \delta < 1 - \delta. \end{aligned}$$

By the probabilistic method, we have proven that there exists some fixed oracle such that A fails with probability $> \delta$. This contradicts the premise that A is a correct agnostic learning algorithm. □

As an interesting aside, note that if we define $\mathbb{C}_\epsilon = \{h_1, h_2, \ldots, h_N\}$, dependent on ϵ, then the agnostic label complexity is $O\left(\log \frac{1}{\epsilon\delta}\right)$ when $\nu = 0$. This is because we can run the realizable learning algorithm to find $f = h_i$, and then sample $\Theta\left(\log \frac{1}{\delta}\right)$ labeled copies of the example $-i$; by observing that they are all labeled $+1$, we effectively *verify* that h_i is at most ϵ-worse than optimal. To make this a correct agnostic algorithm, we can simply be prepared to run A^2 if any of the $\Theta\left(\log \frac{1}{\delta}\right)$ samples of $-i$ are labeled -1 (which they won't be for $\nu = 0$). However, since the disagreement coefficient $\theta = \Theta\left(\frac{1}{\epsilon}\right)$, Theorem 3 implies A^2 does not achieve this improvement. See Appendix B for a similar example.

B. Axis-Aligned Rectangles

Proof Sketch of Theorem 4. To keep things simple, we omit the precise constants. Consider the following algorithm.[3]

0. Sample $\Theta\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ labeled examples from $\mathcal{D}_{XY}$
1. If none of them are positive,
   return the "all negative" concept
2. Else let x be one of the positive examples
3. For $i = 1, 2, \ldots, n$
4. Rejection sample unlabeled set $\mathcal{U}_i$ of size $\Theta\left(\frac{n}{\epsilon\delta}\left(\log\frac{n}{\delta}\right)^2\right)$ from the conditional of $\mathcal{D}$ given $\forall j \neq i, x_j - O\left(\frac{\epsilon\delta}{n\log\frac{1}{\delta}}\right) \leq X_j \leq x_j + O\left(\frac{\epsilon\delta}{n\log\frac{1}{\delta}}\right)$
5. Find $\hat{b}_i = \max\{z_i : z \in \mathcal{U}_i \cup \{x\}, Oracle(z) = +1\}$ by binary search in $\{z_i : z \in \mathcal{U}_i \cup \{x\}, z_i \geq x_i\}$
6. Find $\hat{a}_i = \min\{z_i : z \in \mathcal{U}_i \cup \{x\}, Oracle(z) = +1\}$ by binary search in $\{z_i : z \in \mathcal{U}_i \cup \{x\}, z_i \leq x_i\}$
7. Let $\hat{h} = ((\hat{a}_1, \hat{b}_1), (\hat{a}_2, \hat{b}_2), \ldots, (\hat{a}_n, \hat{b}_n))$
8. Sample $\Theta\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ labeled examples T from $\mathcal{D}_{XY}$
9. If $er_T(\hat{h}) > 0$,
   run A^2 from the start and return its output
10. Else return $\hat{h}$

The correctness of the algorithm in the agnostic setting is clear from examining the three ways to exit the algorithm. First, any oracle with $\mathcal{P}r_{X\sim\mathcal{D}}\{Oracle(X) = +1\} > \epsilon$ will, with probability $\geq 1 - O(\delta)$ have a positive example in the initial $\Theta\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ sample. So if the set has no positives, we can be confident the "all negative" concept has error $\leq \epsilon$. If we return in step 9, we know from Theorem 2 that A^2 will, with probability $1 - O(\delta)$, output a concept with error $\leq \nu + \epsilon$. The remaining possibility is to return in step 10. Any $\hat{h}$ with $er(\hat{h}) > \epsilon$ will, with probability $\geq 1 - O(\delta)$, have $er_T(\hat{h}) > 0$ in step 9. So we can be confident the $\hat{h}$ output in step 10 has $er(\hat{h}) \leq \epsilon$.

To bound the number of label requests, note that the two binary searches we perform for each i (steps 5 and 6) require only $O(\log|\mathcal{U}_i|)$ label requests each, so the entire For loop uses only $O\left(n\log\frac{n}{\epsilon\delta}\right)$ label requests. We additionally have the two labeled sets of size $O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, so if we do not return in step 9, the total number of label requests is at most $O\left(n\log\frac{n}{\epsilon\delta} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$.

It only remains to show that when $\nu = 0$, we do not return in step 9. Let $f = ((a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n))$ be a rectangle with $er(f) = 0$. Note that $er(\hat{h}) \leq \sum_{i=1}^n |a_i - \hat{a}_i| + |b_i - \hat{b}_i|$. For each i, with probability $1 - O(\delta/n)$, none of the initial $\Theta\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ examples w has $w_i \in [a_i, a_i + \gamma] \cup [b_i - \gamma, b_i]$, where $\gamma = O\left(\frac{\epsilon\delta}{n\log\frac{1}{\delta}}\right)$. In particular, if we do not return in step 1, with probability $1 - O(\delta)$, $\forall j, x_j \in [a_j + \gamma, b_j - \gamma]$. Suppose this happens. In particular, this means the oracle's labels for all $z \in \mathcal{U}_i$ are completely determined by whether $a_i \leq z_i \leq b_i$. We can essentially think of this as two "threshold" learning problems for each i: one above x_i and one below x_i. The binary searches find threshold values consistent with each $\mathcal{U}_i$. In particular, by standard passive sample complexity arguments, $|\mathcal{U}_i|$ is sufficient to guarantee with probability $1 - O(\delta/n)$, $|b_i - \hat{b}_i| \leq O\left(\frac{\epsilon\delta}{n\log\frac{1}{\delta}}\right)$ and $|a_i - \hat{a}_i| \leq O\left(\frac{\epsilon\delta}{n\log\frac{1}{\delta}}\right)$. Thus, with probability $1 - O(\delta)$, $er(\hat{h}) \leq O\left(\frac{\epsilon\delta}{\log\frac{1}{\delta}}\right)$. Therefore, the probability $\hat{h}$ makes a mistake on T of size $O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ is at most $O(\delta)$. Otherwise, we have $er_T(\hat{h}) = 0$ in step 9, so we return in step 10. □

Lemma 3. *If $\mathbb{C}$ is the space of axis-aligned rectangles on $[0,1]^n$, and $\mathcal{D}$ is the uniform distribution, then for $\epsilon < \frac{1}{en}$, $\log_2 N(\epsilon/2) \geq n$.*

Proof. Since $N(\epsilon/2)$ is at least the size of any ϵ-separated set, we can prove this lower bound by constructing an ϵ-separated set of size 2^n. In particular, consider the set of all rectangles $((a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n))$ satsifying $\forall i, a_i = 0, b_i \in \left\{1 - \frac{1}{n}, 1\right\}$. There are 2^n such rectangles.

For any two distinct such rectangles $((a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n))$ and $((a_1', b_1'), (a_2', b_2'), \ldots, (a_n', b_n'))$, there is at least one i such that $b_i \neq b_i'$. So the region in which these two disagree contains $\left\{x \in \mathcal{X} : x_i \in \left(1 - \frac{1}{n}, 1\right], \forall j \neq i, x_j \in \left[0, 1 - \frac{1}{n}\right]\right\}$, which has measure $\left(1 - \frac{1}{n}\right)^{n-1}\frac{1}{n} \geq \frac{1}{en} > \epsilon$. □

[3]To keep the algorithm simple, we make little attempt to optimize the number of unlabeled examples. In particular, we could reduce $|\mathcal{U}_i|$ by using a nonzero cutoff in step 9, and could increase the window size in step 4 by using a noise-tolerant active threshold learner in steps 5 and 6.