Active Learning with Covers

July 29, 2014

[AB: As you (Yoav) believed, I have not found previous work on covers that tackles active learning, as active learning appears to not have been a research focus when intensive work on covers was ongoing up to the mid-90s. Though the idea of using covers does come up in recent published work on active learning (by Hanneke), it is relegated to only a brief mention, and it seems like it has not been developed at all before this.]

1 Definitions

We consider a standard statistical learning setting.

Data distribution and labels: Data points ("examples") come from a space $\mathcal{X}$. They are drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X}$.

Each example has a true label $y_i \in \{-1, +1\}$. This is initially unknown by the algorithm but is available upon request. Labels are generated by a conditional distribution $\eta(x) := \Pr(y = +1 \mid x)$.

Hypothesis class: A finite set of classification rules $\mathcal{H} = \{h_1, \ldots, h_n\}$, with each $h_i : \mathcal{X} \mapsto \{-1, +1\}$. These collectively represent all the side information the algorithm uses to make predictions. Each h_i has true error $\operatorname{err}(h_i) = \Pr_{\mathcal{D},\eta}(h_i(x) \neq y)$. On a finite set of labeled examples $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{|S|}, y_{|S|})\}$, the empirical error of h_i is $\widehat{\operatorname{err}}_{=}(h_i) \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbf{1}(h_i(x) \neq y)$.

We denote an optimal hypothesis (there may be multiple) by $h^* \in \arg\min_{h \in \mathcal{H}} \operatorname{err}(h)$. Though much work has been done on the *realizable* case when $\operatorname{err}(h^*) = 0$, we focus on the more general *agnostic* case $\operatorname{err}(h^*) \geq 0$. To this end, define $\mathcal{H}_\alpha = \{h \in \mathcal{H} : \operatorname{err}(h) \leq \operatorname{err}(h^*) + \alpha\}$ for any $\alpha \geq 0$.

Finally, define the *disagreement region* of any $W \subseteq \mathcal{H}$ by $DIS(W) := \{x \in \mathcal{X} : \exists h, h' \in W \; s.t. \; h(x) \neq h'(x)\}$.

Metric embedding of $\mathcal{H}$: The distribution $\mathcal{D}$ induces a (pseudo)metric $\rho(h, h') := \Pr_{x \sim \mathcal{D}}(h(x) \neq h'(x))$ on the hypotheses. With respect to ρ, we can define the closed ball around any $h \in \mathcal{H}$ by $B(h, r) = \{g \in \mathcal{H} : \rho(h, g) \leq r\}$.

Recall that an ϵ-cover of any $W \subseteq \mathcal{H}$ is a subset $\Omega \subseteq W$ such that for any $w \in W$, there exists a $\omega \in \Omega$ such that $\rho(w, \omega) \leq \epsilon$. Write the ϵ-covering number of any $W \subset \mathcal{H}$ w.r.t. the metric ρ as $\mathcal{N}(\epsilon, W, \rho)$, or simply $\mathcal{N}(\epsilon, W)$ when the metric is understood.

Denote a minimum-sized ϵ-cover of W by $\operatorname{cov}(\epsilon, W)$, so that $|\operatorname{cov}(\epsilon, W)| = \mathcal{N}(\epsilon, W)$. (We take the $\operatorname{cov}(\cdot, \cdot)$ operator to be efficient and essentially free to compute compared to the cost of querying labels.) [1]

[1] We gloss over the details of constructing (near-)minimal ϵ-covers for clarity; the discussion of Section 1.1 on unlabeled data addresses some of the related issues.

1.1 Role of Unlabeled Data

We assume that unlabeled examples are available in abundance at essentially no cost. This assumption simplifies the discussion, but we briefly outline its implications here for completeness.

1. **Estimation of Distance Metric**: In particular, this implies that we can determine the distance $\rho(h, h')$ between any pair of classifiers $h, h' \in \mathcal{H}$ to arbitrary accuracy. If needed, sample complexity bounds for this operation can be derived to quantify the amount of unlabeled data used. For instance, if $\mathcal{H}$ has VC dimension d, then classical VC theory implies that all pairwise distances can be estimated uniformly within ϵ using $\mathcal{O}\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$ (unlabeled) examples, w.p. $\geq 1 - \delta$ over the i.i.d. draw of these examples from $\mathcal{D}$.

2. **Infinite $\mathcal{H}$**: Though we assume a finite $\mathcal{H}$, extension to infinite $\mathcal{H}$ is fairly simple. Our finite hypothesis set can be thought of as a high-resolution ϵ-cover of an infinite $\mathcal{H}$ for some very small ϵ.

3. **Constructing a Minimal ϵ-Cover**: Minimal covers of $\mathcal{H}$ at various scales form the core of our approach. They depend on the ability to estimate the disagreement metric ρ as discussed above. In particular, since we take unlabeled data to be free and avoid tractability issues for now, we assume the algorithm can compute (an upper bound on) $\mathcal{N}(\epsilon, W, \rho)$ for any $W \subseteq \mathcal{H}$ and any ϵ.

   A nearly minimal ϵ-cover can be constructed w.p. $\geq 1-\delta$ by first drawing a sample S of $\mathcal{O}\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$ (unlabeled) examples from $\mathcal{D}$ [Haussler 95], and then choosing one classifier from $\mathcal{H}$ for each labeling of S. In this case S is large enough to represent $\mathcal{D}$, of which it is a ($\sim$ minimal) ϵ-cover. [2]

   Alternatively, if $\mathcal{H}$ is finite, constructing an ϵ-net ($\mathcal{O}(\epsilon)$ cover and packing, nearly optimal for both) can be done using a simple greedy iterative algorithm.

   Practically constructing and maintaining these nets has been well-studied, e.g. using cover trees [Beygelzimer, Kakade, Langford 06]. Our analysis contains similar ideas to these and navigating nets [Krauthgamer, Lee 04].

2 Disagreement-Based Active Learning with Covers

[AB: I still need to fill in certain minor parts of the analysis, such as the exact form of a couple of the sample complexity bounds. These parts are marked as such.]

The first active algorithm we analyze is a variant of the well-studied mellow disagreement-based querying strategy first proposed in [Cohn, Atlas, Ladner 94] and often known as "CAL."

2.1 Outline and Motivation

We retain the motivation of CAL: keep track of a version space V of good classifiers which contains h^*, and only query labels of points which are in $DIS(V) := \{x \in \mathcal{X} : \exists h, h' \in V \ s.t. \ h(x) \neq h'(x)\}$.

The algorithm we study (Algorithm 1) differs somewhat from the original CAL, in that it handles the agnostic case [3] and only polls a representative cover of V to decide whether to query labels. However, in a similar spirit to CAL and other disagreement-based algorithms, our algorithm estimates the excess error of every hypothesis in V, in order to update V.

This raises one issue in particular: the examples that the algorithm leaves unlabeled can only be useful in estimating excess errors to within Ω(resolution of the cover). Since the cover resolution must be made

[2] This approach is basically the same as in [Buescher, Kumar 96], and they develop it in detail; for our purposes, the relevant part of their analysis comprises Problem 12.14 of [Devroye, Győrfi, Lugosi 96].

[3] Most of the analysis therefore resembles that of [Dasgupta, Hsu, Monteleoni 07] and sequels. This line of work is summarized in [Hanneke 14, Ch. 5].

progressively finer as the algorithm progresses, this means that examples which are left unlabeled in earlier iterations cannot be used to estimate error in later iterations.

Our crude solution to this issue is to throw away these earlier examples, which of course does not the learning process. This is a central problem in designing an aggressive cover-based active learning algorithm, and should be addressed better in future work.

2.2 Preliminaries

Before analyzing Algorithm 1, we will define some placeholder quantities.

We use standard uniform convergence bounds. These state that for any $W \subseteq \mathcal{H}$, if S is a set of m i.i.d. (labeled) examples drawn with $(\mathcal{D}, \eta)$, then w.p. $\geq 1-\delta$, for all $h \in W$,

$$\left|\widehat{\mathrm{err}}_S(h) - \mathrm{err}(h)\right| \leq F(W, \delta, m) \in o(1) \tag{1}$$

Here $F(\cdot,\cdot,\cdot)$ is a function of the complexity of W.

2.2.1 Notes on F and δ

F can be explicitly written by using a classical empirical Rademacher complexity result, that w.h.p.

$$\max_{h \in W} \left|\widehat{\mathrm{err}}_S(h) - \mathrm{err}(h)\right| \leq 2\hat{\mathcal{R}}_S(W) + 3\sqrt{\frac{\log(1/\delta)}{m}}$$

and the empirical Rademacher complexity can then be bounded by the log covering number, or VC dimension, or just $\sim \log|\mathcal{H}|$ (Massart lemma). One bound that may be useful to us is Pollard's discretization bound $\hat{\mathcal{R}}_S(W) \lesssim \epsilon + \sqrt{\frac{2\log \mathcal{N}(\epsilon, W)}{m}}$ for any ϵ (actually the empirical covering number, hence "$\lesssim$"). See the discussion in Section 2.4 for more concreteness.

[AB: I will fill in the details and references for $F(\cdot,\cdot,\cdot)$, depending on how we want to quantify complexity. The algorithm actually needs to compute and use this bound directly; any of the above bounds seems fine for this purpose.]

[AB: I also neglect to properly allocate the failure probabilities δ for now. The allocations I am glossing over are only of two kinds: union bounds over a small number of events in each epoch, and a union bound over the countable set of epochs. Both are straightforward and don't significantly affect the final results.]

2.3 Analysis: Generalization Error

Let S_i be the set of all m_i examples sampled in epoch i, along with their labels. S_i will only be a tool for the analysis, since the algorithm does not query all labels in S_i.

Most of our analysis focuses on an arbitrary epoch i; for this reason, we omit the subscripts from L_i, m_i, S_i, V_i when the context is clear.

Define S to be the set of all (unlabeled) examples sampled by the algorithm, so that $|S| = m$ after each iteration.

The general idea of the analysis is to uniformly estimate the excess errors of all "good" hypotheses h using S, and in particular the labeled set L.

In other words, for any $h \in V$,

$$\widehat{\mathrm{err}}_S(h) - \arg\min_{g \in V} \widehat{\mathrm{err}}_S(g) \tag{2}$$

is used as a surrogate for $\mathrm{err}(h) - \mathrm{err}(h^*)$. Now since $|S| = m$, (2) can be decomposed into two disjoint parts dealing with L and $S \setminus L$, for any $g \in V$:

$$\widehat{\mathrm{err}}_S(h) - \widehat{\mathrm{err}}_S(g) = \frac{|L|}{m}\left(\widehat{\mathrm{err}}_L(h) - \widehat{\mathrm{err}}_L(g)\right) + \frac{m-|L|}{m}\left(\widehat{\mathrm{err}}_{S\setminus L}(h) - \widehat{\mathrm{err}}_{S\setminus L}(g)\right) \tag{3}$$

Algorithm 1 Agnostic CAL with Covers

Given: Desired accuracy γ, failure probability δ
Define: cover scales $\epsilon_k \leftarrow 2^{-k}$ for $k \in \mathbb{N}$
Initialize: $V_0 \leftarrow \mathcal{H}$; $W_0 \leftarrow \mathrm{cov}(\epsilon_0, V_0)$; epoch counter $i \leftarrow 0$; for $i = 1, 2, \ldots$: $L_i \leftarrow \emptyset$, unlabeled data counters $m_i \leftarrow 0$
while $i < \log_2(\gamma^{-1})$ **do**
  $m_i \leftarrow m_i + 1$
  Sample an unlabeled example $X^i_{m_i} \sim \mathcal{D}$
  if $X^i_{m_i} \in DIS(W_i)$ **then**
    Query label $Y^i_{m_i}$ of $X^i_{m_i}$, and set $L_i \leftarrow L_i \cup \{(X^i_{m_i}, Y^i_{m_i})\}$
  end if
  if m_i is large enough that $3\sqrt{\dfrac{\log\left(\frac{4\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right)}{m_i}} + 2F\left(V_i, \dfrac{\delta}{2}, m_i\right) \le 4\epsilon_i$ **then**
    $$V_{i+1} \leftarrow \left\{ h \in V_i : \frac{|L|}{m}\left(\widehat{\mathrm{err}}_{L_i}(h) - \min_{g \in V_i} \widehat{\mathrm{err}}_{L_i}(g)\right) \le 2F\left(V_i, \frac{\delta}{2}, m_i\right) + \frac{\epsilon_i}{\Pr\left(x \notin DIS(W_i)\right)} + \sqrt{\frac{\log\left(\frac{2\mathcal{N}(\epsilon_i, V_i)}{\delta}\right)}{2m_i}} \right\}$$
    $W_{i+1} \leftarrow \mathrm{cov}(\epsilon_{i+1}, V_{i+1})$
    $i \leftarrow i + 1$
  end if
end while
return Any $h \in V_i$

Clearly, the L-dependent term is known exactly from the labeled dataset. To deal with the $(S \setminus L)$-dependent term, we must characterize the behavior of the excess risk on the unqueried points, when $x \notin DIS(W_i)$. The following simple lemma concerns this.

Lemma 1. *Take any $g \in V_i$ and any $h \in W_i$. Then*

$$\Pr\left(g(x) \neq h(x) \mid x \notin DIS(W_i)\right) \le \frac{\epsilon_i}{\Pr\left(x \notin DIS(W_i)\right)}$$

Proof. Suppose $g_C \in \arg\min_{c \in W_i} \rho(g, c)$. Since W_i is an ϵ_i-cover,

$$\begin{aligned}\epsilon_i &\ge \Pr\left(g_C(x) \neq g(x)\right) \ge \Pr\left(g_C(x) \neq g(x) \mid x \notin DIS(W_i)\right) \Pr\left(x \notin DIS(W_i)\right) \\ &= \Pr\left(h(x) \neq g(x) \mid x \notin DIS(W_i)\right) \Pr\left(x \notin DIS(W_i)\right)\end{aligned}$$

where the last line is because $h(x) = g_C(x)$ whenever $x \notin DIS(W_i)$. ∎

Lemma 1 can be used immediately to control the second term of (3), leading to the following result.

Lemma 2. *Take any $g \in V_i$ and any $h \in W_i$, and suppose S is a set of m (labeled) examples drawn i.i.d. from $\mathcal{D}, \eta$. Also suppose $\Pr\left(x \notin DIS(W_i)\right) \ge 2\epsilon_i$. Then w.p. $\ge 1 - \delta$,*

$$\left|\left(\widehat{err}_S(g) - \widehat{err}_S(h)\right) - \frac{|L|}{m}\left(\widehat{err}_L(g) - \widehat{err}_L(h)\right)\right| \le \frac{\epsilon_i}{\Pr\left(x \notin DIS(W_i)\right)} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Proof. Define $p := \Pr\left(g(x) = y, h(x) \neq y \mid x \notin DIS(W_i)\right)$ and $q := \Pr\left(g(x) \neq h(x) \mid x \notin DIS(W_i)\right)$. Each example in $S \setminus L$ is drawn i.i.d., and contributes one of three values $\{-\frac{1}{m}, 0, \frac{1}{m}\}$ to $\widehat{\mathrm{err}}_{S \setminus L}(g) - \widehat{\mathrm{err}}_{S \setminus L}(h)$,

with respective probabilities $\{p, 1-q, q-p\}$. Therefore each example contributes a value with magnitude $\leq \frac{1}{m}\text{Ber}(q)$ to $\widehat{\text{err}}_{S\setminus L}(g) - \widehat{\text{err}}_{S\setminus L}(h)$.

So $\left|\frac{m-|L|}{m}\left(\widehat{\text{err}}_{S\setminus L}(g) - \widehat{\text{err}}_{S\setminus L}(h)\right)\right| = \frac{1}{m}X$ where $X \sim \text{Bin}(m-|L|, q)$. Combining this with (3),

$$\left|\left(\widehat{\text{err}}_S(g) - \widehat{\text{err}}_S(h)\right) - \frac{|L|}{m}\left(\widehat{\text{err}}_L(g) - \widehat{\text{err}}_L(h)\right)\right| \leq \frac{1}{m}X$$

Recall from Lemma 1 that $q \leq \frac{\epsilon_i}{\Pr(x \notin DIS(W_i))} \leq \frac{1}{2}$. Therefore, applying a Chernoff bound on X gives the result. ■

Lemma 3. *Suppose $h^* \in V_i$. Then w.p. $\geq 1-\delta$, for all $h \in W_i$,*

$$\widehat{err}_S(h^*) - \widehat{err}_S(h) \leq 2F(V_i, \delta, m)$$

Proof. Using the optimality of h^* and then (1),

$$\begin{aligned}\widehat{\text{err}}_S(h^*) - \widehat{\text{err}}_S(g) &= \left(\widehat{\text{err}}_S(h^*) - \text{err}(h^*)\right) + \left(\text{err}(h^*) - \text{err}(g)\right) + \left(\text{err}(g) - \widehat{\text{err}}_S(g)\right) \\ &\leq \left(\widehat{\text{err}}_S(h^*) - \text{err}(h^*)\right) + \left(\text{err}(g) - \widehat{\text{err}}_S(g)\right) \leq 2F(V_i, \delta, m)\end{aligned}$$

■

Our accounting for the unlabeled examples in S establishes that V always contains h^*.

Lemma 4. *With probability $\geq 1-\delta$, for all epochs i, $h^* \in V_i$.*

Proof. This is proved by induction on i. The base case is clear because $h^* \in \mathcal{H} = V_0$.

For the inductive case, assume $h^* \in V_i$. Then by taking a union bound of Lemma 2 over all $h \in W_i$, we have that w.p. $\geq 1 - \frac{\delta}{2}$, for all $h \in W_i$,

$$\frac{|L|}{m}\left(\widehat{\text{err}}_L(h^*) - \widehat{\text{err}}_L(h)\right) \leq \widehat{\text{err}}_S(h^*) - \widehat{\text{err}}_S(h) + \frac{\epsilon_i}{\Pr(x \notin DIS(W_i))} + \sqrt{\frac{\log\left(\frac{2\mathcal{N}(\epsilon_i, V_i)}{\delta}\right)}{2m}} \tag{4}$$

Substituting Lemma 3 into (4) and taking a union bound, we get that w.p. $\geq 1-\delta$,

$$\frac{|L|}{m}\left(\widehat{\text{err}}_L(h^*) - \min_{h \in W_i}\widehat{\text{err}}_L(h)\right) \leq 2F\left(V_i, \frac{\delta}{2}, m\right) + \frac{\epsilon_i}{\Pr(x \notin DIS(W_i))} + \sqrt{\frac{\log\left(\frac{2\mathcal{N}(\epsilon_i, V_i)}{\delta}\right)}{2m}}$$

The left-hand side of this inequality is $\geq \frac{|L|}{m}\left(\widehat{\text{err}}_L(h^*) - \min_{h \in V_i}\widehat{\text{err}}_L(h)\right)$ because $W_i \subseteq V_i$. This implies that $h^* \in V_{i+1}$, completing the induction. ■

Since the version space V_i always contains h^* and our selective sampling controls the relative risk of any hypotheses in V_i, we can derive a generalization error bound on any hypothesis in V_i (and thereby on the hypothesis returned by the algorithm).

Theorem 5. *For all epochs i, w.h.p.*

$$\max_{h \in V_{i+1}} err(h) - err(h^*) \leq \frac{8\epsilon_i}{\Pr(x \notin DIS(W_i))}$$

Proof. We prove the result for an arbitrary epoch i.

It suffices to prove that w.h.p.

$$\max_{h \in V_{i+1}} \operatorname{err}(h) - \operatorname{err}(h^*) \leq \frac{4\epsilon_i}{\Pr(x \notin DIS(W_i))} + 3\sqrt{\frac{\log\left(\frac{4\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right)}{m}} + 2F\left(V_i, \frac{\delta}{2}, m\right) \tag{5}$$

because of the definition of $m := m_i$ in the algorithm (Line 10).

First define $w^* \in \arg\min_{g \in W_i} \rho(g, h^*)$. Since $h^* \in V_i$ by Lemma 4, $\rho(w^*, h^*) \leq \epsilon_i$, so we have by (1) that w.p. $\geq 1 - \delta$, for all $h \in W_{i+1} \subseteq V_{i+1} \subseteq V_i$ simultaneously,

$$\begin{aligned}
\operatorname{err}(h) - \operatorname{err}(w^*) &= (\operatorname{err}(h) - \widehat{\operatorname{err}}_S(h)) + (\widehat{\operatorname{err}}_S(h) - \widehat{\operatorname{err}}_S(w^*)) + (\widehat{\operatorname{err}}_S(w^*) - \operatorname{err}(w^*)) \\
&\leq (\widehat{\operatorname{err}}_S(h) - \widehat{\operatorname{err}}_S(w^*)) + \sqrt{\frac{2\log\left(\frac{(\mathcal{N}(\epsilon_i, V_i) + \mathcal{N}(\epsilon_{i+1}, V_{i+1}))}{\delta}\right)}{m}}
\end{aligned}$$

Combining this with Lemma 2 and a union bound over all $h \in W_{i+1}$, w.p. $\geq 1 - \delta$, $\forall h \in W_{i+1}$,

$$\operatorname{err}(h) - \operatorname{err}(w^*) \leq \frac{|L|}{m}(\widehat{\operatorname{err}}_L(h) - \widehat{\operatorname{err}}_L(w^*)) + \frac{\epsilon_i}{\Pr(x \notin DIS(W_i))} + \sqrt{\frac{5\log\left(\frac{2(\mathcal{N}(\epsilon_i, V_i) + \mathcal{N}(\epsilon_{i+1}, V_{i+1}))}{\delta}\right)}{m}} \tag{6}$$

Now we examine the term $\frac{|L|}{m}(\widehat{\operatorname{err}}_L(h) - \widehat{\operatorname{err}}_L(w^*))$. If $w^* \in V_{i+1}$, then since $h \in V_{i+1}$, by definition of the V_{i+1} update (Line 11 of the algorithm), we have

$$\frac{|L|}{m}(\widehat{\operatorname{err}}_L(h) - \widehat{\operatorname{err}}_L(w^*)) \leq 2F\left(V_i, \frac{\delta}{2}, m\right) + \frac{\epsilon_i}{\Pr(x \notin DIS(W_i))} + \sqrt{\frac{\log\left(\frac{2\mathcal{N}(\epsilon_i, V_i)}{\delta}\right)}{2m}} \tag{7}$$

Alternatively, if $w^* \notin V_{i+1}$, since $w^* \in V_i$ and $h \in V_{i+1}$, we have (again by definition of the V_{i+1} update) that $\frac{|L|}{m}(\widehat{\operatorname{err}}_L(w^*) - \widehat{\operatorname{err}}_L(h)) \geq 0$. Combining this with (7) and substituting into (6),

$$\begin{aligned}
\max_{h \in W_{i+1}} \operatorname{err}(h) - \operatorname{err}(w^*) &\leq \frac{2\epsilon_i}{\Pr(x \notin DIS(W_i))} + 3\sqrt{\frac{\log\left(\frac{2(\mathcal{N}(\epsilon_i, V_i) + \mathcal{N}(\epsilon_{i+1}, V_{i+1}))}{\delta}\right)}{m}} + 2F\left(V_i, \frac{\delta}{2}, m\right) \\
&\leq \frac{2\epsilon_i}{\Pr(x \notin DIS(W_i))} + 3\sqrt{\frac{\log\left(\frac{4\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right)}{m}} + 2F\left(V_i, \frac{\delta}{2}, m\right)
\end{aligned} \tag{8}$$

where the last line is because $\mathcal{N}(\epsilon, V)$ is monotonic decreasing in ϵ and increasing in V, $\epsilon_i \geq \epsilon_{i+1}$, and $V_{i+1} \subseteq V_i$.

The result (5) now follows by combining (8) with the relation

$$\begin{aligned}
\max_{h \in V_{i+1}} \operatorname{err}(h) - \operatorname{err}(h^*) &\overset{(a)}{\leq} \max_{h \in W_{i+1}} \operatorname{err}(h) - \operatorname{err}(h^*) + \epsilon_{i+1} \overset{(b)}{\leq} \max_{h \in W_{i+1}} \operatorname{err}(h) - \operatorname{err}(w^*) + \epsilon_{i+1} + \epsilon_i \\
&\leq \max_{h \in W_{i+1}} \operatorname{err}(h) - \operatorname{err}(w^*) + 2\epsilon_i
\end{aligned}$$

where (a) is because W_{i+1} is an ϵ_{i+1}-cover of V_{i+1} and (b) is by the definition of w^*. ■

The excess error in Theorem 5, $\frac{8\epsilon_i}{\Pr(x \notin DIS(W_i))}$, is $\mathcal{O}(\epsilon_i)$ in many cases. For an example, recall the definition of the disagreement coefficient $\theta = \sup_{\epsilon} \frac{\Pr(DIS(B(h^*, \epsilon)))}{\epsilon}$, widely used to derive guarantees for

disagreement-based active learning algorithms. Then we have the following when θ is finite (proof omitted here). [4]

Corollary 6. *If $\mathcal{H}$ has disagreement coefficient $\theta < \infty$, then for all epochs $i \geq 1 + \log_2(\theta)$, w.h.p.*

$$\max_{h \in V_{i+1}} err(h) - err(h^*) \leq 16\epsilon_i$$

This is essentially as good as this cover-based algorithm can hope for, since the resolution of the cover in epoch i is also $\mathcal{O}(\epsilon_i)$.

2.4 Analysis: Label Complexity

Having derived a generalization error guarantee, we now show that the algorithm does not query too many labels. For classes $\mathcal{H}$ with a finite disagreement coefficient, we can prove a label complexity bound.

In order to quantify label complexity, we must explicitly show a bound on the number of unlabeled examples m_i used. This can only be done if we make $F(\cdot,\cdot,\cdot)$ concrete. So for our purposes in this section, we follow the discussion in Section 2.2.1, which gives a family of possible F depending on metric entropy, by Pollard's discretization result. We choose F based on the epoch i:

$$F(V_i, \delta, m_i) :\approx 2\epsilon_{i+1} + 2\sqrt{\frac{2\log \mathcal{N}(\epsilon_{i+1}, V_i)}{m_i}} + 3\sqrt{\frac{\log(1/\delta)}{m_i}} \leq \epsilon_i + 6\sqrt{\frac{\log\left(\frac{\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right)}{m_i}} \tag{9}$$

(**Note**: The "$\approx$" is because Pollard's discretization result bounding $\hat{\mathcal{R}}_S(V_i)$ only deals with the covering number using the empirical metric calculated with S, not the metric ρ. So a further approximation result is required to state a precise bound. Due to this, all results in this subsection are only correct up to constants.)

Now we can turn the condition that implicitly defines m_i (Line 10 of the algorithm) into an explicit result on m_i.

Lemma 7. *For any epoch i,*

$$m_i = \left\lceil \frac{225}{4\epsilon_i^2} \log\left(\frac{4\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right) \right\rceil$$

Proof. This is a simple consequence of substituting (9) into the condition defining m_i (Line 10 of the algorithm), and then solving for m_i. ■

Theorem 8. *If $\mathcal{H}$ has disagreement coefficient $\theta < \infty$, then define $i_0 := 1 + \log_2(\theta)$. Suppose $\max_i \mathcal{N}(\epsilon_{i+1}, V_i) \leq 2^d$ for some d. Then w.h.p. for any target excess error γ that the algorithm is given, the algorithm queries*

$$\mathcal{O}\left(\theta\left(\frac{err(h^*)+\gamma}{\gamma^2}\right)\log\left(\frac{1}{\gamma}\right)\left(d + \log\left(\frac{1}{\delta}\right)\right)\right)$$

labels from epoch $i_0 + 1$ onwards.

Proof. Since $\theta < \infty$, the argument is similar to standard label complexity analyses ([Dasgupta, Hsu, Monteleoni 07] and sequels), but we give a self-contained proof here.

Note that $\mathcal{H}_\alpha \subseteq B(h^*, 2\text{err}(h^*) + \alpha)$ by the triangle inequality.

By Corollary 6, for $i > i_0$, we have $W_i \subseteq V_i \subseteq \mathcal{H}_{16\epsilon_{i-1}} \subseteq B(h^*, 2\text{err}(h^*) + 16\epsilon_{i-1})$, so by definition of θ,

$$\Pr(x \in DIS(W_i)) \leq \Pr(x \in DIS(B(h^*, 2\text{err}(h^*) + 16\epsilon_{i-1}))) \leq \theta(2\text{err}(h^*) + 16\epsilon_{i-1}) \tag{10}$$

[4] A sketch: Suppose $\mathcal{H}$ has finite disagreement coefficient θ, and let $\mathcal{H}_\alpha = \{h \in \mathcal{H} : \text{err}(h) \leq \text{err}(h^*) + \alpha\}$. Roughly speaking, since $W_i \subseteq V_i \subseteq \mathcal{H}_{\epsilon_i}$, $\Pr(x \notin DIS(W_i)) \geq \Pr(x \notin DIS(V_i)) = 1 - \Pr(x \in DIS(\mathcal{H}_{\epsilon_i})) \geq 1 - \theta\epsilon_i$, so the result is shown. The inclusion $V_i \subseteq \mathcal{H}_{\epsilon_i}$ does not strictly follow from Theorem 5, but can be proven by induction on i if $\theta < \infty$.

On examining the algorithm, we have for $i > i_0$ that $|L_i| = \sum_{j=1}^{m_i} \mathbf{1}(X_j^i \in DIS(W_i))$. This is a sum of independent Bernoulli variables, each with success probability bounded by (10); so applying a Chernoff bound, union bounds, and (10), w.h.p. for all $i > i_0$,

$$\begin{aligned}
|L_i| &= \sum_{j=1}^{m_i} \mathbf{1}(X_j^i \in DIS(W_i)) \leq \sqrt{m_i \log(1/\delta)} + m_i \Pr\left(x \in DIS(W_i)\right) \\
&\leq \sqrt{m_i \log(2/\delta)} + m_i \theta(2\mathrm{err}\left(h^*\right) + 16\epsilon_{i-1}) \leq \theta(2\mathrm{err}\left(h^*\right) + 16\epsilon_{i-1}) \frac{225}{2\epsilon_i^2} \log\left(\frac{4\mathcal{N}(\epsilon_{i+1}, V_i)}{\delta}\right) \\
&= \mathcal{O}\left(\theta\left(\frac{\mathrm{err}\left(h^*\right) + \gamma}{\gamma^2}\right)\left(d + \log\left(\frac{1}{\delta}\right)\right)\right)
\end{aligned}$$

The result follows by noting that there are $\lesssim \log(1/\gamma)$ epochs. ■

We have intentionally deferred taking a supremum over epochs until this final point of the analysis; it is intuitive that the complexity of the algorithm depends on a measure of complexity (covering number) that is *local* to V_i. If at any point there are many approximately error-minimizing hypotheses well-separated in ρ, they might be difficult to detect with any active learning algorithm that searches the hypothesis space. [AB: This suggests the possibility of a lower bound in terms of local complexities, which I will think about.] [AB: TBD: Include comparison with known distribution-dependent lower bounds for active learning.]

2.5 Generalizing the Disagreement Coefficient

[AB:

- When θ is finite, the region of unanimity of the hypotheses is large and only takes one hypothesis to cover. What upper bound can be derived on $\mathcal{N}(\epsilon_{i+1}, \mathcal{H}_{\epsilon_i})$?
- Revisit the "Smooth Relative Regret Approximations" of [Ailon, Begleiter, Ezra 12], as they directly generalize the disagreement coefficient in a way that appears similar to our strategy of taking a cover and sampling from its disagreement region. Specifically, is one epoch of our algorithm equivalent to taking their relative regret approximation?

]

3 QBC with Covers

We next analyze the only general-purpose aggressive active learning algorithm in the theoretical literature, the query-by-committee algorithm.

3.1 Generalizing the Splitting Index?