Agnostic Active Learning

Maria-Florina Balcan NINAMF@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3891

Alina Beygelzimer BEYGEL@US.IBM.COM
IBM T. J. Watson Research Center, Hawthorne, NY 10532

John Langford JL@TTI-C.ORG
Toyota Technological Institute at Chicago, Chicago, IL 60637

Abstract

We state and analyze the first active learning algorithm which works in the presence of arbitrary forms of noise. The algorithm, A^2 (for *A*gnostic *A*ctive), relies only upon the assumption that the samples are drawn *i.i.d.* from a fixed distribution. We show that A^2 achieves an exponential improvement (i.e., requires only $O\left(\ln \frac{1}{\epsilon}\right)$ samples to find an ϵ-optimal classifier) over the usual sample complexity of supervised learning, for several settings considered before in the realizable case. These include learning threshold classifiers and learning homogeneous linear separators with respect to an input distribution which is uniform over the unit sphere.

1. Introduction

What distinguishes active learning from the more typical batch learning is that the algorithm initially sees only the unlabeled portion of a pool of examples drawn from some underlying distribution. The algorithm can then pay for the label of any example in the pool, and the hope is that a good classifier can be learned with significantly fewer labels by actively directing the queries to informative examples. There is a significant practical interest in minimizing the number of labeled examples in settings where there is no shortage of unlabeled data but labels are expensive.

Most active learning strategies are *noise seeking* on many natural learning problems. In particular, the process of trying to find an optimal separation between one class and another often involves label queries for examples with a large conditional noise rate. Thus the most informative examples are also the ones that are typically the most noise-prone.

Consider the active learning algorithm which searches for the optimal threshold on an interval using binary search. This example is often used to demonstrate the potential of active learning in the noise-free case when there is a perfect threshold separating the classes (Cohen et al., 1994). Binary search needs $O(\ln \frac{1}{\epsilon})$ labeled examples to learn a threshold with error less than ϵ, while learning passively requires $O\left(\frac{1}{\epsilon}\right)$ labels. A fundamental drawback of this algorithm is that a small amount of adversarial noise can force the algorithm to behave badly. Is this extreme brittleness to small amounts of noise essential? Can we still achieve an exponential decrease in sample complexity? Can we avoid making assumptions about the mechanism producing noise? These are the questions addressed here.

Previous Work on Active Learning There has been substantial work on active learning under additional assumptions.

For example, the Query by Committee analysis (Freund et al., 1993) assumes realizability (i.e., there exists a perfect classifier in a known set) and a correct Bayesian prior on the set of hypotheses. Recent work by Dasgupta (2005) has identified sufficient and semi-necessary conditions for active learning given only the additional realizability assumption. (Dasgupta et al., 2005; Dasgupta, 2004) also assume realizability. If there exists a perfect separator in our class, we can use any informative querying strategy to direct the learning process without the need to worry about the distribution it induces—any inconsistent hypothesis can be eliminated based on a *single* query, regardless of which distribution this query comes from. In the agnostic case, however, a hypothesis that performs badly on the query distribution may well be the optimal hypothesis with respect to the input distribution. This is the main challenge in agnostic active learning that

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

is not present in the non-agnostic case.

Burnashev and Zigangirov (1974) allow noise, but require a correct Bayesian prior on threshold functions. Other assumptions include specific noise models. For example, Castro et al. (2005) have analyzed active learning in a setting with a constant noise rate everywhere.

The *membership-query* setting (Angluin, 1998; Angluin, 2001; Bshouty & Eiron, 2002; Jackson, 1997) is similar to active learning considered here except that no unlabeled data is given. Instead, the learning algorithm is allowed to query examples of its own choice. This is problematic in several applications because natural oracles, such as hired humans, have difficulty labeling synthetic examples (Baum & Lang, 1992). Ulam's Problem (quoted in (Czyzowicz et al., 1989)), where the goal is find a distinguished element in a set by asking subset membership queries, is also related. The quantity of interest is the smallest number of such queries required to find the element, given a bound on the number of queries that can be answered incorrectly. But both types of results do not apply here since an active learning strategy can only buy labels of the examples it observes. For example, a membership query algorithm can be used to quickly find a separating hyperplane in a high-dimensional space. An active learning algorithm can not do so when the data distribution does not support queries close to the decision boundary.

When Active Learning Can Help It is important to keep in mind that the speedups achievable with active learning depend on the match between the distribution over example-label pairs and the hypothesis class, and therefore on the target hypothesis in the class. There are simple examples where active learning does not help at all, even if there is no noise (see, for example, (Dasgupta, 2005)). Obviously, all such lower bounds apply in our setting as well.

It is also important to note that we cannot hope to achieve speedups when the noise rate η is large, due to a lower bound of $\Omega(\frac{\eta^2}{\epsilon^2})$ on the sample complexity of any active learner (Kaariainen, 2005).

Summary of Results In section 3, we present an *A*gnostic *A*ctive learning algorithm, A^2. The only assumption we rely upon is that samples are drawn *i.i.d.* from some underlying distribution. In particular, we make no assumptions about the mechanism producing noise (e.g., class/target misfit, fundamental randomization, adversarial situations). As far as we know, this is the first result of this form.

Section 3.1 proves that A^2 is correct, and section 3.2 shows that A^2 is never harmful, i.e., it never requires significantly more samples than batch learning. Section 4 shows the potential of A^2 by establishing exponential speedups in several settings previously analyzed without noise. In particular, we show that we can achieve exponential speedups for the simple case of learning threshold functions; the result holds for arbitrary distributions provided that the noise rate is sufficiently small with respect to the desired error ϵ. We also show that our algorithm achieves an exponential improvement if the hypothesis class consists of homogeneous (through the origin) linear separators and the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$. The last example has been the most encouraging theoretical result so far in the realizable case (Dasgupta et al., 2005). The A^2 analysis also achieves an almost contradictory property: for some sets of classifiers, we can choose an ϵ-optimal classifier with fewer samples than are needed to estimate the error rate of the chosen classifier with precision ϵ.

2. Preliminaries

Let X be an instance space and $Y = \{-1, 1\}$ be the set of possible labels. Let H be the hypothesis class, a set of functions mapping from X to Y. We assume there is a distribution D over instances in X, and that the instances are labeled by a possibly randomized oracle O. The *error rate* of a hypothesis h with respect to a distribution P over $X \times Y$ is defined as $\mathrm{err}_P(h) = \mathbf{Pr}_{x,y\sim P}[h(x) \neq y]$. Let $\eta = \min_{h \in H}(\mathrm{err}_{D,O}(h))$ denote the minimum error rate of any hypothesis in H with respect to the distribution (D, O) induced by D and the labeling oracle O. The goal is to find a hypothesis $h \in H$ with $\mathrm{err}_{D,O}(h)$ within ϵ of η, where ϵ is some target error.

The algorithm A^2 relies on a subroutine, which computes a lower bound $\mathrm{LB}(S, h, \delta)$ and an upper bound $\mathrm{UB}(S, h, \delta)$ on the true error rate $\mathrm{err}_P(h)$ of h by using a sample S of examples drawn *i.i.d.* from P. Each of these bounds must hold for all h simultaneously with probability at least $1 - \delta$. The subroutine is formally defined below.

Definition 1 *A subroutine for computing $LB(S, h, \delta)$ and $UB(S, h, \delta)$ is said to be* legal *if for all distributions P over $X \times Y$, and for all $m \in \mathbb{N}$,*

$$LB(S, h, \delta) \leq \mathrm{err}_P(h) \leq UB(S, h, \delta)$$

holds for all $h \in H$ simultaneously, with probability $1 - \delta$ over the draw of S according to P^m.

Examples of such subroutines are the VC bound (Vapnik & Chervonenkis, 1971) and the Occam Razor bound (Blumer et al., 1987).

3. The A^2 Agnostic Active Learner

At a high level, A^2 can be viewed as a robust version of the selective sampling algorithm of Cohen et al. (1994). Selective sampling is a sequential process that keeps track of two spaces—the current *version space* H_i, defined as the set of hypotheses in H consistent with all labels revealed so far, and the current *region of uncertainty* R_i, defined as the set of all $x \in X$, for which there exists a pair of hypotheses in H_i that disagrees on x. In round i, the algorithm picks a random unlabeled example from R_i and queries it, eliminating all hypotheses in H_i inconsistent with the received label. The algorithm then eliminates those $x \in R_i$ on which all surviving hypotheses agree, and recurses. This process fundamentally relies on the assumption that there exists a consistent hypothesis in H. In the agnostic case, we cannot eliminate a hypothesis based on its disagreement with a single example. We need to be more conservative, or we risk eliminating best hypotheses in the class.

A formal specification of A^2 is given in Algorithm 1. Let H_i be the set of hypotheses still under consideration by A^2 in round i. If all hypotheses in H_i agree on some region of the instance space, this region can be safely eliminated. To help us keep track of progress in decreasing the region of uncertainty, define $\textsc{Disagree}_D(H_i)$ as the probability that there exists a pair of hypotheses in H_i that disagrees on a random example drawn from D:

$$\textsc{Disagree}_D(H_i) = \mathbf{Pr}_{x \sim D}[\exists h_1, h_2 \in G : h_1(x) \neq h_2(x)].$$

Hence $\textsc{Disagree}_D(H_i)$ is the volume of the current region of uncertainty with respect to D.

Let D_i be the distribution D restricted to the current region of uncertainty. Formally, $D_i = D(x \mid \exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x))$. In round i, A^2 samples a set of examples S_i from D_i, O, and uses it to compute upper and lower bounds for all hypotheses in H_i. It then eliminates all hypotheses whose lower bound is greater than the minimum upper bound. Figure 3.1 shows the algorithm in action for the case when the data lie in the $[0, 1]$ interval on the real line, and H is the set of thresholding functions. The horizontal axis denotes both the instance space and the hypothesis space, superimposed. The vertical axis shows the error rates. Round i completes when S_i is large enough to eliminate at least half of the current region of uncertainty. Since we eliminate only those examples on which the surviving hypotheses agree, an optimal hypothesis in H_i with respect to D_i remains an optimal hypothesis in H_{i+1} with respect to D_{i+1}. Since each round i cuts $\textsc{Disagree}_D(H_i)$ down by half, the number of rounds is bounded by $\log \frac{1}{\epsilon}$. Sections 4 gives examples of distributions and hypothesis classes for which A^2 requires only a small number of labeled examples to transition between rounds, yielding an exponential improvement in sample complexity.

When evaluating bounds during the course of Algorithm 1, we choose a schedule of δ according to the following rule: we evaluate bound k with confidence $\delta_k = \frac{\delta}{k(k+1)}$, for $k \geq 1$.

Algorithm 1 A^2 (allowed error rate ϵ, sampling oracle for D, labeling oracle O, hypothesis class H)

```
Set $i = 1$, $D_i = D$, $H_i = H$, $S_i = \emptyset$, and $k = 1$.
while $\textsc{Disagree}_D(H_i)(\min_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k) - \min_{h \in H_i} \mathrm{LB}(S_i, h, \delta_k)) > \epsilon$
  1. set $S_i = \emptyset$, $H'_i = H_i$, $k \leftarrow k + 1$.
  2. while $\textsc{Disagree}_D(H'_i) \geq \frac{1}{2}\textsc{Disagree}_D(H_i)$
     (a) if $\textsc{Disagree}_D(H'_i)(\min_{h \in H'_i} \mathrm{UB}(S_i, h, \delta_k) - \min_{h \in H'_i} \mathrm{LB}(S_i, h, \delta_k)) \leq \epsilon$
     (b)     return $h = \mathrm{argmin}_{h \in H'_i} \mathrm{UB}(S_i, h, \delta_k)$.
     (c) else
         i. $S'_i$ = Rejection sample $2|S_i| + 1$ samples $x$ from $D$ satisfying:
               $\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x)$
         ii. $S_i \leftarrow S_i \cup \{(x, O(x)) : x \in S'_i\}$; $k \leftarrow k + 1$;
         iii. $H'_i = \{h \in H_i : \mathrm{LB}(S_i, h, \delta_k,) \leq \min_{h' \in H_i} \mathrm{UB}(S_i, h', \delta_k)\}$, $k \leftarrow k + 1$.
         end if
     end while
  3. $H_{i+1} \leftarrow H'_i$;
     $D_{i+1} \leftarrow D_i$ conditioned on the disagreement $\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x)$;
     $i \leftarrow i + 1$.
end while
Return $h = \mathrm{argmin}_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k)$.
```

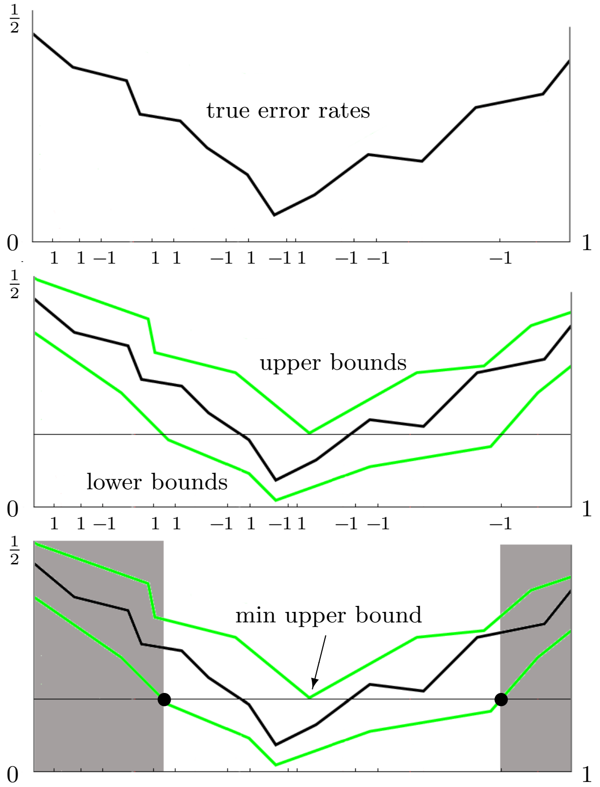

Figure 3.1. A^2 in action: Sampling, Bounding, Eliminating.

3.1. Correctness

Theorem 3.1 (Correctness) *For all H, for all (D, O), for all valid subroutines for computing UB and LB, with probability $1-\delta$, A^2 returns an ϵ-optimal hypothesis or does not terminate.*

Note 1 *For most "reasonable" subroutines for computing UB and LB, A^2 terminates with probability at least $1 - \delta$. For more discussion and a proof of this fact see Section 3.2.*

Proof: We first prove that all bound evaluations are valid simultaneously with probability at least $1-\delta$, and then show that the procedure produces an ϵ-optimal hypothesis upon termination.

To prove the first claim, notice that the samples on which each bound is evaluated are drawn *i.i.d.* from some distribution over $X \times Y$. This can be verified by noting that the distribution D_i used in round i is precisely that given by drawing x from the underlying distribution D conditioned on the disagreement $\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x)$, and then labeling according to the oracle O. The k-th bound evaluation fails with probability at most $\frac{\delta}{k(k+1)}$. By the Union bound, the probability that any bound fails is less then the sum of the probabilities of individual bound failures. This sum is bounded by $\sum_{k=1}^{\infty} \frac{\delta}{k(k+1)} = \delta$.

To prove the second claim, notice first that since every bound evaluation is correct, we can be certain that step 2(c)(iii) never eliminates a hypothesis that has minimum error rate with respect D. Let us now introduce the following notation. For a hypothesis $h \in H$ and $G \subseteq H$ define:

$$e_{D,G,O}(h) = \mathbf{Pr}_{x,y \sim D,O|\exists h_1,h_2 \in G: h_1(x) \neq h_2(x)}[h(x) \neq y],$$

$$f_{D,G,O}(h) = \mathbf{Pr}_{x,y \sim D,O|\forall h_1,h_2 \in G: h_1(x) = h_2(x)}[h(x) \neq y].$$

Notice that $e_{D,G,O}(h)$ is in fact $\mathrm{err}_{D_G,O}(h)$, where D_G is D conditioned on the disagreement $\exists h_1, h_2 \in G : h_1(x) \neq h_2(x)$. Moreover, given any $G \subseteq H$, we can decompose the error rate of every hypothesis h into two parts as follows:

$\mathrm{err}_{D,O}(h) = e_{D,G,O}(h) \cdot \textsc{Disagree}_D(G) + f_{D,G,O}(h) \cdot (1 - \textsc{Disagree}_D(G)) = \mathrm{err}_{D_G,O}(h) \cdot \textsc{Disagree}_D(G) + f_{D,G,O}(h) \cdot (1 - \textsc{Disagree}_D(G))$.

Notice that the only term that varies with $h \in G$ in the above decomposition, is $e_{D,G,O}(h)$. Consequently, if we want to find an ϵ-optimal hypothesis we need only bound the error rate of $\mathrm{err}_{D_G,O}(h) \cdot \textsc{Disagree}_D(G)$ to precision ϵ. But this is exactly what the negation of the main while-loop guard does, and this is also the condition we use in Step 2(a) of the algorithm. In other words, upon termination we have $\textsc{Disagree}_D(H_i)(\min_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k) - \min_{h \in H_i} \mathrm{LB}(S_i, h, \delta_k)) \leq \epsilon$, which proves the desired result. ∎

3.2. Fallback Analysis

This section shows that A^2 is never much worse than a standard batch, bound-based algorithm[1] in terms of the number of samples required in order to learn, assuming that UB and LB are "sane".

Define the sample complexity $m(\varepsilon, \delta, H)$ required by a batch algorithm that uses a subroutine for computing $\mathrm{LB}(S, h, \delta)$ and $\mathrm{UB}(S, h, \delta)$ as the minimum number of samples m such that for all $S \in X^m$, we have $|\mathrm{UB}(S, h, \delta) - \mathrm{LB}(S, h, \delta)| \leq \varepsilon$ for all $h \in H$.

We use the following bound on $m(\epsilon, \delta, H)$ stated as Theorem A.1 in Appendix A:

$$m(\epsilon, \delta, H) = \frac{64}{\epsilon^2}\left(2V \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right)$$

Here V is the VC-dimension of H. Assume that $m(2\epsilon, \delta, H) \leq \frac{m(\epsilon,\delta,H)}{2}$, and also that the function m

[1] A standard example of a bound-based learning algorithm is Empirical Risk Minimization (ERM).

is monotonically increasing in $1/\delta$. These conditions are satisfied by many subroutines for computing UB and LB, including those based on the VC-bound (Vapnik & Chervonenkis, 1971) and the Occam's Razor bound (Blumer et al., 1987).

Theorem 3.2 *For all H, for all (D, O), for all UB and LB satisfying the assumption above, the algorithm A^2 makes at most $2m(\epsilon, \delta', H)$ calls to the oracle O, where $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and $N(\epsilon, \delta, H)$ satisfies: $N(\epsilon, \delta, H) \geq \ln\frac{1}{\epsilon}\ln m(\epsilon, \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}, H)$. Here $m(\epsilon, \delta, H)$ is the sample complexity of UB and LB.*

Proof: Let $\delta_k = \frac{\delta}{k(k+1)}$ be the confidence parameter used in the k-th application of the subroutine for computing UB and LB. We will determine an upper bound $N(\epsilon, \delta, H)$ on the number of bound evaluations throughout the life of the algorithm. This will imply that the confidence parameter δ_k will always be greater than $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$.

Consider $i = 1$. If condition 2 of Algorithm A^2 is repeatedly satisfied then after labeling $m(\epsilon, \delta', H)$ examples from D_1 we have that, uniformly, for all hypotheses $h \in H_1$, $|\mathrm{UB}(S_1, h, \delta') - \mathrm{LB}(S_1, h, \delta')| \leq \epsilon$. Note that in these conditions we safely halt. Notice also that the number of bound evaluations during this process is at most $\ln m(\epsilon, \delta', H)$.

On the other hand, if loop (2) ever completes and i increases, then it is enough to have uniformly for all $h \in H_2$, $|\mathrm{UB}(S_2, h, \delta') - \mathrm{LB}(S_2, h, \delta')| \leq 2\epsilon$. (This follows from the exit conditions we use in the outer while-loop and in Step 2(a) of A^2.) Clearly, in order to uniformly have that for all hypotheses $h \in H_2$ their true upper and lower bounds are within 2ϵ from each other, we only need $m(2\epsilon, \delta', H) \leq \frac{m(\epsilon,\delta',H)}{2}$ labeled examples from D_2 and the number of bounds evaluations in round $i = 2$ is at most $\ln m(\epsilon, \delta', H)$.

In general, in round i it is enough to have uniformly for all $h \in H_i$, $|\mathrm{UB}(S_i, h, \delta') - \mathrm{LB}(S_i, h, \delta')| \leq 2^{i-1}\epsilon$, and in order to obtain this we only need $m(2^{i-1}\epsilon, \delta', H) \leq \frac{m(\epsilon,\delta',H)}{2^{i-1}}$ labeled examples from D_i. Also the number of bounds evaluations in round i is at most $\ln m(\epsilon, \delta', H)$.

Since the number of rounds is bounded by $\ln\frac{1}{\epsilon}$, it follows that the maximum number of bound evaluation throughout the life of the algorithm is at most $\ln\frac{1}{\epsilon}\ln m(\epsilon, \delta', H)$. This implies that in order to determine an upper bound $N(\epsilon, \delta, H)$ we just need to find a solution of the following inequality: $N(\epsilon, \delta, H) \geq \ln\frac{1}{\epsilon}\ln m(\epsilon, \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}, H)$.

Finally, adding up the number of calls to the oracle in all rounds, we get that the number of calls to the oracle throughout the life of the algorithm is at most $2m(\epsilon, \delta', H)$. $\blacksquare$

Let V denote the VC-dimension of H, and let $m(\epsilon, \delta, H)$ be the number of examples required by the ERM algorithm. As we state in Theorem A.1 in Appendix A a classic bound on $m(\epsilon, \delta, H)$ is $m(\epsilon, \delta, H) = \frac{64}{\epsilon^2}\left(2V\ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right)$. Then using Theorem 3.2, we can show the following corollary.

Corollary 3.3 *For all hypothesis classes H of VC-dimension V, for all distributions (D, O) over $X \times Y$, the algorithm A^2 requires at most $\tilde{O}\left(\frac{1}{\epsilon^2}(V\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})\right)$ labeled examples drawn* i.i.d. *from (D, O).*[2]

Proof: We use the form of $m(\epsilon, \delta, H)$ and Theorem 3.2 to upper bound $N = N(\epsilon, \delta, H)$. It is enough to find the smallest N satisfying $N \geq \ln\left(\frac{1}{\epsilon}\right)\ln\left(\frac{64}{\epsilon^2}\left(2V\ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4N^2}{\delta}\right)\right)\right)$. Using the inequality $\ln a \leq ab - \ln b - 1$ for all $a, b > 0$ and some simple algebraic manipulations, we get the desired upper bound on $N(\epsilon, \delta, H)$. The result then follows from Theorem 3.2. $\blacksquare$

4. Active Learning Speedups

In this section, we show that it is possible to achieve exponential sample complexity improvements even with arbitrary noise for some sets of classifiers.

4.1. Learning Threshold Functions

We begin by analyzing the simple class of threshold functions. As mentioned in the introduction, it turns out that even for this simple class of functions exponential reduction in sample complexity is *not* achievable when the noise rate η is large (Kaariainen, 2005). Therefore we prove two results: one shows an exponential sample complexity improvement when the noise rate is small, while the other simply shows a slower improvement when the noise rate is large. In the extreme where the noise rate is $1/2$, there is no improvement.

Theorem 4.1 *Let H be the set of thresholds on an interval with LB and UB the VC bound. For all distributions (D, O), for any $\epsilon < \frac{1}{2}$ and $\eta < \frac{\epsilon}{16}$, the algorithm A^2 makes $O\left(\ln\left(\frac{1}{\epsilon}\right)\ln\left(\frac{\ln\left(\frac{1}{\epsilon\delta}\right)}{\delta}\right)\right)$ calls to the oracle O on examples drawn* i.i.d. *from D, with probability $1 - \delta$.*

[2] Here and in the rest of the paper, the $\tilde{O}(\cdot)$ notation is used to hide factors logarithmic in the factors present explicitly.

Proof: Each execution of loop 2 decreases $\textsc{Disagree}_D(H_i)$ by at least a factor of 2, implying that the number of executions is bounded by $\log \frac{1}{\epsilon}$. Consequently, we are done if only $O\left(\ln \frac{1}{\delta'}\right)$ labeled samples are required per loop.[3]

Let h^* be any minimum error rate hypothesis. For $h_1, h_2 \in H_i$, let $d_i(h_1, h_2)$ be the probability that h_1 and h_2 predict differently on a random example drawn according to D_i, i.e., $d_i(h_1, h_2) = \mathbf{Pr}_{x \sim D_i}(h_1(x) \neq h_2(x))$.

Consider $i \geq 1$. Let $[lower_i, upper_i]$ be the support of D_i. Note that for any hypothesis $h \in H_i$ we have $\mathrm{err}_{D_i,O}(h) \geq d_i(h, h^*) - \mathrm{err}_{D_i,O}(h^*)$. We also clearly have $\mathrm{err}_{D_i,O}(h^*) \leq \eta/Z_i$, where $Z_i = \mathbf{Pr}_{x \sim D}(x \in [lower_i, upper_i])$. (So Z_i is a shorthand for $\textsc{Disagree}_D(H_i)$.)

Now notice that at least a $\frac{1}{2}$-fraction (measured with respect to D_i) of thresholds in H_i satisfy $d_i(h, h^*) \geq \frac{1}{4}$, and these thresholds are located at the "ends" of the interval $[lower_i, upper_i]$. Formally, assume first that both $d_i(h^*, lower_i) \geq \frac{1}{4}$ and $d_i(h^*, upper_i) \geq \frac{1}{4}$, then let l_i and u_i be the hypotheses to the left and to the right of h^*, respectively, that satisfy $d_i(h^*, l_i) = \frac{1}{4}$ and $d_i(h^*, u_i) = \frac{1}{4}$. We clearly have that all $h \in [lower_i, l_i] \cup [u_i, upper_i]$ satisfy $d_i(h^*, h) \geq \frac{1}{4}$ and moreover $\mathbf{Pr}_{x \sim D_i}(x \in [lower_i, l_i] \cup [u_i, upper_i]) \geq \frac{1}{2}$. Now assume without loss of generality that $d_i(h^*, lower_i) \leq \frac{1}{4}$. Let u_i be the hypothesis to the right of h^* with $d_i(h, upper_i) = \frac{1}{2}$. Then we clearly have that all $h \in [u_i, upper_i]$ satisfy $d_i(h^*, h) \geq \frac{1}{4}$ and moreover $\mathbf{Pr}_{x \sim D_i}(x \in [u_i, upper_i]) \geq \frac{1}{2}$.

Using the VC bound, we get that with probability $1 - \delta'$ if

$$|S_i| = O\left(\frac{\ln \frac{1}{\delta'}}{\left(\frac{1}{8} - \frac{\eta}{Z_i}\right)^2}\right),$$

then uniformly for all hypotheses $h \in H_i$, we have $|\mathrm{UB}(S_i, h, \delta) - \mathrm{LB}(S_i, h, \delta)| \leq \frac{1}{8} - \frac{\eta}{Z_i}$.

Consider a hypothesis $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. For any such h we have $\mathrm{err}_{D_i,O}(h) \geq d_i(h, h^*) - \mathrm{err}_{D_i,O}(h^*) \geq \frac{1}{4} - \frac{\eta}{Z_i}$, and so $\mathrm{LB}(S_i, h, \delta) \geq \frac{1}{4} - \frac{\eta}{Z_i} - (\frac{1}{8} - \frac{\eta}{Z_i}) = \frac{1}{8}$. On the other hand, $\mathrm{err}_{D_i,O}(h^*) \leq \frac{\eta}{Z_i}$, and so $\mathrm{UB}(S_i, h^*, \delta) \leq \frac{\eta}{Z_i} + \frac{1}{8} - \frac{\eta}{Z_i} = \frac{1}{8}$. Thus A^2 eliminates all $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. But that means we have $\textsc{Disagree}_D(H_i') \leq \frac{1}{2}\textsc{Disagree}_D(H_i)$, terminating round i. [4]

Finally notice that A^2 makes $O\left(\ln\left(\frac{1}{\delta'}\right) \ln\left(\frac{1}{\epsilon}\right)\right)$ calls to the oracle, where $\delta' = \frac{\delta}{N^2(\epsilon, \delta, H)}$ and $N(\epsilon, \delta, H)$ is an upper bound on the number of bound evaluations throughout the life of the algorithm. We clearly have that the number of bound evaluations required in round i is $O\left(\frac{\ln \frac{1}{\delta'}}{\left(\frac{1}{8} - \frac{\eta}{Z_i}\right)^2}\right)$. This implies that the number of bound evaluations throughout the life of the algorithm $N(\epsilon, \delta, H)$ should satisfy $c \ln\left(\frac{N^2(\epsilon, \delta, H)}{\delta}\right) \ln\left(\frac{1}{\epsilon}\right) \leq N(\epsilon, \delta, H)$, for some constant c. Solving this inequality, we get the desired result. ∎

Theorem 4.2 *Let H be the set of thresholds on an interval with LB and UB the VC bound. Suppose that $\epsilon < \frac{1}{2}$ and $\eta > \epsilon$. For all D, with probability $1 - \delta$, the algorithm A^2 will require at most $\tilde{O}\left(\frac{\eta^2 \ln \frac{1}{\delta}}{\epsilon^2}\right)$ labeled samples.*

Proof Sketch: The proof is similar to the previous proof. Theorem 4.1 implies that loop (2) will complete $\log \frac{1}{\eta} - 4$ times. At this point, the noise becomes sufficient so that the algorithm may only halt via the return step in loop (2). In this case, we have $\textsc{Disagree}_D(H) = \Theta(\eta)$ implying that the number of samples required is $\tilde{O}\left(\frac{\eta^2 \ln \frac{1}{\delta}}{\epsilon^2}\right)$. ∎

Note that Theorem 4.2 asymptotically matches a lower bound of Kaariainen (Kaariainen, 2005).

4.2. Linear Separators under the Uniform Distribution

A commonly analyzed case for which active learning is known to give exponential savings in the number of labeled examples is when the data is drawn uniformly from the unit sphere in $\mathbb{R}^d$, and the labels are consistent with a linear separator going through the origin. We show that A^2 also gives exponential savings in this case, in the presence of noise.

Let $X = \{x \in \mathbb{R}^d : \|x\| = 1\}$, the unit sphere in $\mathbb{R}^d$. Assume that D is uniform over X, and let H be the class of linear separators through the origin. Any $h \in H$ is a homogeneous hyperplane represented by a unit vector $w \in X$ with the classification rule $h(x) = \mathrm{sign}(w \cdot x)$. The distance between two hypotheses u and v in H with respect to a distribution D (i.e., the probability that they predict differently on a random example drawn from D) is given by $d_D(u, v) =$

[3] Notice that the difference $(\min_{h \in H_i} \mathrm{UB}(S_i, h, D_i, O) - \min_{h \in H_i} \mathrm{LB}(S_i, h, D_i, O))$ appearing in Step 2(a) is always constant.

[4] The assumption in the theorem statement can clearly be weakened to $\eta < \frac{\epsilon}{(8+\Delta)\sqrt{d}}$ for any constant $\Delta > 0$.

$\frac{\arccos(u\cdot v)}{\pi}$. Finally, let $\theta(u,v) = \arccos(u\cdot v)$. Thus $d_D(u,v) = \frac{\theta(u,v)}{\pi}$.

Theorem 4.3 *Let X, H, and D be as defined above, and let LB and UB be the VC bound. Then for any $0 < \epsilon < \frac{1}{2}$, $0 < \eta < \frac{\epsilon}{16\sqrt{d}}$, and $\delta > 0$, with probability $1-\delta$, A^2 requires*

$$O\left(d\left(d\ln d + \ln\frac{1}{\delta'}\right)\ln\frac{1}{\epsilon}\right)$$

calls to the labeling oracle, where $\delta' = \frac{\delta}{N^2(\epsilon,\delta,H)}$ and $N(\epsilon,\delta,H) = O\left(\ln\frac{1}{\epsilon}\left(d^2\ln d + \ln\frac{d\ln\frac{1}{\epsilon}}{\delta}\right)\right)$.

Proof: Let $w^* \in H$ be a hypothesis with the minimum error rate η. Denote the region of uncertainty in round i by R_i. Thus $\mathbf{Pr}_{x\sim D}[x \in R_i] = \text{DISAGREE}_D(H_i)$.

Consider round i of A^2. Initially $i = 1$, corresponding to $H_i = H$, $D_i = D$, and $\text{DISAGREE}_D(H_i) = 1$.

Theorem A.1 says that it suffices to query the oracle on a set S of $O(d^2\ln d + d\ln\frac{1}{\delta'})$ examples from D_i to guarantee, with probability $1-\delta'$, that for all $w \in H_i$,

$$|\text{err}_{D_i,O}(w) - \widehat{\text{err}}_{D_i,O}(w)| < \frac{1}{2}\left(\frac{1}{8\sqrt{d}} - \frac{\eta}{r_i}\right),$$

where r_i is a shorthand for $\text{DISAGREE}_D(H_i)$. (By assumption, $\eta < \frac{\epsilon}{16\sqrt{d}}$. We also have $\text{DISAGREE}_D(H_i) \geq \epsilon$. Thus the precision above is at least $\frac{1}{16\sqrt{d}}$.)[5] This implies that $\text{UB}(S,w,\delta') - \text{err}_{D_i,O}(w) < \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i}$, and $\text{err}_{D_i,O}(w) - \text{LB}(S,w,\delta') < \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i}$. Consider any $w \in H_i$ with $d_{D_i}(w,w^*) \geq \frac{1}{4\sqrt{d}}$. For any such w, we have $\text{err}_{D_i,O}(w) \geq \frac{1}{4\sqrt{d}} - \frac{\eta}{r_i}$, and so

$$\text{LB}(S,w,\delta') > \frac{1}{4\sqrt{d}} - \frac{\eta}{r_i} - \frac{1}{8\sqrt{d}} + \frac{\eta}{r_i} = \frac{1}{8\sqrt{d}}.$$

But we also know that $\text{err}_{D_i,O}(w^*) = \frac{\eta}{r_i}$, and thus $\text{UB}(S,w^*,\delta') < \frac{\eta}{r_i} + \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i} = \frac{1}{8\sqrt{d}}$, so A^2 will eliminate w in step 2(c)(iii).

Thus round i eliminates all hypotheses $w \in H_i$ with $d_{D_i}(w,w^*) \geq \frac{1}{4\sqrt{d}}$. Since all hypotheses in H_i agree on every $x \notin R_i$, we have

$$d_{D_i}(w,w^*) = \frac{1}{r_i}d_D(w,w^*) = \frac{\theta(w,w^*)}{\pi r_i}.$$

Thus round i eliminates all hypotheses $w \in H_i$ with $\theta(w,w^*) \geq \frac{\pi r_i}{4\sqrt{d}}$. But since $2\theta/\pi \leq \sin\theta$, for $\theta \in (0,\frac{\pi}{2}]$, it certainly eliminates all w with $\sin\theta(w,w^*) \geq \frac{r_i}{2\sqrt{d}}$.

Consider any $x \in R_{i+1}$ and the value $|w^*\cdot x| = \cos\theta(w^*,x)$. There must exist a hypothesis $w \in H_{i+1}$ that disagrees with w^* on x; otherwise x would not be in R_{i+1}. But then we must have $\cos\theta(w^*,x) \leq \cos(\frac{\pi}{2} - \theta(w,w^*)) = \sin\theta(w,w^*) < \frac{r_i}{2\sqrt{d}}$, where the last inequality is due to the fact that A^2 eliminated all w with $\sin\theta(w,w^*) \geq \frac{r_i}{2\sqrt{d}}$. Thus any $x \in R_{i+1}$ must satisfy $|w^*\cdot x| < \frac{r_i}{2\sqrt{d}}$.

Using the fact that $\mathbf{Pr}(A\,|\,B) = \frac{\mathbf{Pr}(AB)}{\mathbf{Pr}(B)} \leq \frac{\mathbf{Pr}(A)}{\mathbf{Pr}(B)}$ for any A and B, we can write

$$\mathbf{Pr}_{x\sim D_i}[x \in R_{i+1}] \leq \mathbf{Pr}_{x\sim D_i}\left[|w\cdot x| \leq \frac{r_i}{2\sqrt{d}}\right]$$
$$\leq \frac{\mathbf{Pr}_{x\sim D}\left[|w\cdot x| \leq \frac{r_i}{2\sqrt{d}}\right]}{\mathbf{Pr}_{x\sim D}[x \in R_i]} \leq \frac{r_i}{2r_i} = \frac{1}{2},$$

where the second inequality follows from Lemma A.2. Thus $\text{DISAGREE}_D(H_{i+1}) \leq \frac{1}{2}\text{DISAGREE}_D(H_i)$, as desired.

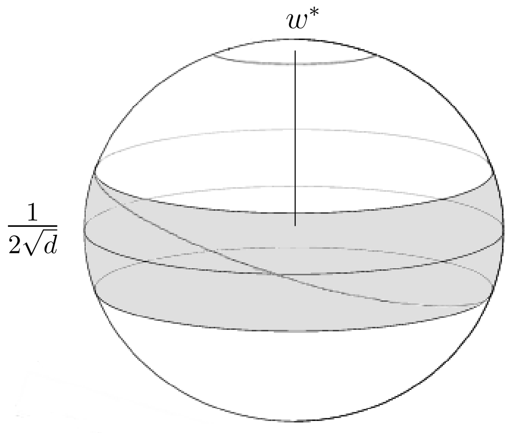

Figure 4.1. The region of uncertainty after the first iteration (schematic).

To finish the argument, it suffices to notice that since every round cuts $\text{DISAGREE}_D(H_i)$ at least in half, the total number of rounds is bounded by $\log\frac{1}{\epsilon}$. Notice also that A^2 makes $O\left(d^2\ln d + d\ln\frac{1}{\delta'}\right)\ln\left(\frac{1}{\epsilon}\right)$ calls to the oracle, where $\delta' = \frac{\delta}{N^2(\epsilon,\delta,H)}$ and $N(\epsilon,\delta,H)$ is an upper bound on the number of bound evaluations throughout the life of the algorithm. We clearly have that the number of bound evaluations required in round i is $O\left(d^2\ln d + d\ln\frac{1}{\delta'}\right)$. This implies that the number of bound evaluations throughout the life of the algorithm $N(\epsilon,\delta,H)$ should satisfy $c\left(d^2\ln d + d\ln\left(\frac{N^2(\epsilon,\delta,H)}{\delta}\right)\right)\ln\left(\frac{1}{\epsilon}\right) \leq N(\epsilon,\delta,H)$ for some constant c. Solving this inequality, we get the desired result. ∎

[5] The assumption in the theorem statement can clearly be weakened to $\eta < \frac{\epsilon}{(8+\Delta)\sqrt{d}}$ for any constant $\Delta > 0$.

For comparison, the query complexity of the Perceptron-based active learner of Dasgupta et al. (2005), is $O(d \ln \frac{1}{\epsilon\delta}(\ln \frac{d}{\delta} + \ln\ln \frac{1}{\epsilon}))$, for the same H, X, and D, but only for the realizable case when $\eta = 0$.

5. Discussion and Open Questions

The results here should be regarded as a first-case proof-of-possibility. A^2 suggests a number of interesting open questions:

1. On what other (hypothesis spaces, distribution) pairs can we observe exponential speedups? Is there an algorithm that is more sample efficient than A^2? Does A^2 always achieve speedups comparable to the ones achieved by the selective sampling algorithm (Cohen et al., 1994), but in the presence of (limited) noise?

2. Are there concept classes for which A^2 (or some other algorithm) can be made computationally efficient? Checking for disagreement amongst all remaining classifiers can be very computationally intensive.

3. What conditions are sufficient and necessary for active learning to succeed in the agnostic case?

References

Angluin, D. (1998). Queries and concept learning. *Machine Learning*, *2*, 319–342.

Angluin, D. (2001). Queries revisited. *Proceedings of the 12th International Conference on Learning Theory.*

Anthony, M., & Bartlett, P. (1999). *Neural Network Learning: Theoretical Foundations.* Cambridge University Press.

Baum, E., & Lang, K. (1992). Query learning can work poorly when a human oracle is used. *International Joint Conference on Neural Networks.*

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1987). Occam's razor. *Information Processing Letters*, *24*, 377–380.

Bshouty, N. H., & Eiron, N. (2002). Learning monotone dnf from a teacher that almost does not answer membership queries. *Journal of Machine Learning Research*, *3*, 49–57.

Burnashev, M., & Zigangirov, K. (1974). An interval estimation problem for controlled observations. *Problems in Information Transmission*, *10*, 223–231.

Castro, R., Willett, R., & Nowak, R. (2005). Fast rates in regression via active learning. *University of Wisconsin Technical Report ECE-05-03.*

Cohen, D., Atlas, L., & Ladner, R. (1994). Improving generalzation with active learning. *Machine Learning*, *15(2)*, 201–221.

Czyzowicz, J., Mundici, D., & Pelc, A. (1989). Ulam's searching game with lies. *Journal of Combinatorial Theory, Series A*, *52*, 62–76.

Dasgupta, S. (2004). Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems (NIPS).*

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

Dasgupta, S., Kalai, A., & Monteleoni, C. (2005). Analysis of perceptron-based active learning. *COLT.*

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1993). Information, prediction, and query by comittee. *Neural Information Processing Systems.*

Jackson, J. (1997). An efficient membership-query algorithm for learning dnf with respect to the uniform distribution. *Journal of Computer and System Sciences*, *55(3).*

Kaariainen, M. (2005). On active learning in the nonrealizable case. *NIPS Workshop on Foundations of Active Learning.*

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16(2)*, 264–280.

A. Appendix

We use the following standard Sample Complexity bound from Anthony and Bartlett (1999).

Theorem A.1 *Suppose that H is a set of functions from X to $\{-1, 1\}$ with finite VC-dimension $V \geq 1$. Let D be an arbitrary, but fixed probability distribution over $X \times \{-1, 1\}$. For any ϵ, $\delta > 0$, if we draw a sample from D of size*

$$m(\epsilon, \delta, V) = \frac{64}{\epsilon^2}\left(2V \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right),$$

then with probability at least $1 - \delta$, we have $|err(h) - \widehat{err}(h)| \leq \epsilon$ for all $h \in H$.

In section 4.2 we make use of the following lemma. For a proof see, for example, Dasgupta et al. (2005).

Lemma A.2 *For any fixed unit vector w and any $0 < \gamma \leq 1$,*

$$\frac{\gamma}{4} \leq \mathbf{Pr}_x\left[|w \cdot x| \leq \frac{\gamma}{\sqrt{d}}\right] \leq \gamma,$$

where x is drawn uniformly from the unit sphere.