Agnostic Active Learning ⋆

Maria-Florina Balcan

Carnegie Mellon University, Pittsburgh, PA 15213

Alina Beygelzimer

IBM T. J. Watson Research Center, Hawthorne, NY 10532

John Langford

Yahoo! Research, New York, NY 10011

Abstract

We state and analyze the first active learning algorithm that finds an ϵ-optimal hypothesis in any hypothesis class, when the underlying distribution has arbitrary forms of noise. The algorithm, A^2 (for *A*gnostic *A*ctive), relies only upon the assumption that it has access to a stream of unlabeled examples 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.

Key words: Active Learning, Agnostic Setting, Sample Complexity, Linear Separators.

1 Introduction

Traditionally, machine learning has focused on the problem of learning a task from labeled examples only. In many applications, however, labeling

⋆ A preliminary version of this paper appeared in the 23rd International Conference on Machine Learning, ICML 2006.

Email addresses: ninamf@cs.cmu.edu (Maria-Florina Balcan), beygel@us.ibm.com (Alina Beygelzimer), jl@yahoo-inc.com (John Langford).

is expensive while unlabeled data is usually ample. This observation motivated substantial work on properly using unlabeled data to benefit learning [4,10,11,30,34,28,33], and there are many examples showing that unlabeled data can significantly help [9,32].

There are two main frameworks for incorporating unlabeled data into the learning process. The first framework is *semi-supervised learning* [18], where in addition to a set of labeled examples, the learning algorithm can also use a (usually larger) set of unlabeled examples drawn at random from the same underlying data distribution. In this setting, unlabeled data becomes useful under additional assumptions and beliefs about the learning problem. For example, transductive SVM learning [28] assumes that the target function cuts through low density regions of the space, while co-training [11] assumes that the target should be self-consistent in some way. Unlabeled data is potentially useful in this setting because it allows one to reduce the search space to a set which is a-priori reasonable with respect to the underlying distribution.

The second setting, which is the main focus of this paper, is *active learning* [19,22]. Here the learning algorithm is allowed to draw random unlabeled examples from the underlying distribution and ask for the labels of any of these examples. The hope is that a good classifier can be learned with significantly fewer labels by actively directing the queries to informative examples.

As in passive supervised learning, but unlike in semi-supervised learning, the only prior belief about the learning problem here is that the target function (or a good approximation of it) belongs to a given concept class. For some concept classes such as thresholds on the line, one can achieve an exponential improvement over the usual sample complexity of supervised learning, under no additional assumptions about the learning problem [19,22]. In general, the speedups achievable in active learning depend on the match between the data distribution and the hypothesis class, and therefore on the target hypothesis in the class. The most noteworthy non-trivial example of improvement is the case of homogeneous (i.e., through the origin) linear separators, when the data is linearly separable and distributed uniformly over the unit sphere [25,24,22]. There are also simple examples where active learning does not help at all, even in the realizable case [22].

Most of the previous work on active learning has focused on the realizable case. In fact, many of the existing active learning strategies are *noise seeking* on natural learning problems, because the process of actively finding an optimal separation between one class and another often involves label queries for examples close to the decision boundary, and such examples often used a large conditional noise rate (e.g., due to a mismatch between the hypothesis class and the data distribution). Thus the most informative examples are also the ones that are typically the most noise-prone.

Consider an 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 [19]. 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 an exponential decrease in sample complexity be achieved? Can assumptions about the mechanism producing noise be avoided? 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 [25] assumes realizability (i.e., existence of a perfect classifier in a known set), and a correct Bayesian prior on the set of hypotheses. Dasgupta [22] has identified sufficient conditions (which are also necessary against an adversarially chosen distribution) for active learning given only the additional realizability assumption. There are several other papers that assume only realizability [21,24]. If there exists a perfect separator amongst hypotheses, any informative querying strategy can 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 is not present in the non-agnostic case. Burnashev and Zigangirov [15] allow noise, but require a correct Bayesian prior on threshold functions. Some papers require specific noise models such as a constant noise rate everywhere [17] or Tsybakov noise conditions [5,16].

The *membership-query* setting [1,2,14,27] 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 [8]. Ulam's Problem (quoted in [20]), 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.

Our Contributions This paper presents the first *a*gnostic *a*ctive learning algorithm, A^2. The only necessary assumption is that the algorithm has access to a stream of examples drawn *i.i.d.* from some fixed distribution. No additional assumptions are made about the mechanism producing noise (e.g., class/target misfit, fundamental randomization, adversarial situations). The main contribution of this paper is to prove the feasibility of agnostic active learning.

Two comments are in order:

(1) We define the *noise rate* of a hypothesis class H with respect to a fixed distribution D as the minimum error rate of any hypothesis in H on D (see section 2 for a formal definition). Note that for the special case of so called *label noise* (where a coin of constant bias is used to determine whether any particular example is mislabeled with respect to the best hypothesis) these definitions coincide.
(2) We regard unlabeled data as being free so as to focus exclusively on the question of whether or not agnostic active learning is possible at all. Substantial follow-up work to this paper has successfully optimized unlabeled data usage to be on the same order as passive learning [23].

A^2 is provably correct (for any $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, it outputs an ϵ-optimal hypothesis with probability at least $1 - \delta$) and it is never harmful (it never requires significantly more labeled examples than batch learning). A^2 provides exponential sample complexity reductions in several settings previously analyzed without noise or with known noise conditions. This includes learning threshold functions with small noise with respect to ϵ and hypothesis classes consisting of homogeneous (through the origin) linear separators with the data 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 [24].

The A^2 analysis achieves an almost contradictory property: for some sets of classifiers, an ϵ-optimal classifier can be output with fewer labeled examples than are needed to estimate the error rate of the chosen classifier with precision ϵ from random examples only.

Lower Bounds 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. Thus one should expect the results to be distribution-dependent. There are simple examples where active learning does not help at all in the model analyzed in this paper, even if there is no noise [22]. These lower bounds essentially result from an "aliasing" effect and they are unavoidable in the setting we analyze in this paper (where we bound the number of

queries an algorithm makes before it *can prove* it has found a good function). [1]

In the noisy situation, the target function itself can be very simple (e.g., a threshold function), but if the error rate is very close to 1/2 in a sizeable interval near the threshold, then no active learning procedure can significantly outperform passive learning. In particular, in the pure agnostic setting one *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 [29]. However, under specific noise models (such as a constant noise rate everywhere [17] or Tsybakov noise conditions [5,16]) and for specific classes, one can still show significant improvement over supervised learning.

Structure of This Paper Preliminaries and notation are covered in Section 2, then A^2 is presented in Section 3. Section Section 3.1 proves that A^2 is correct and Section 3.2 proves it is never harmful (i.e., it never requires significantly more samples than batch learning). Threshold functions such as $f_t(x) = sign(x - t)$ and homogeneous linear separators under the uniform distribution over the unit sphere are analyzed in Section 4. Conclusions, a discussion of subsequent work, and open questions are covered in Section 5.

2 Preliminaries

We consider a binary agnostic learning problem specified as follows. 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 oracle O can be thought of as taking an unlabeled example x in, choosing a biased coin based on x, then flipping it find the label -1 or 1. The *error rate* of a hypothesis h with respect to a distribution P over $X \times Y$ is defined as $\mathrm{err}_P(h) = \mathbf{Pr}_{x,y\sim P}[h(x) \neq y]$. The error rate $\mathrm{err}_P(h)$ is not generally known since P is unknown, however the empirical version $\widehat{\mathrm{err}}_P(h) = \mathbf{Pr}_{x,y\sim S}[h(x) \neq y] = \frac{1}{S}\sum_{x,y\in S} I(h(x) \neq y)$ is computable based upon an observed sample set S.

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 an ϵ-optimal hypothesis, i.e. a hypothesis $h \in H$ with

[1] In recent work, Balcan et. al [6,7] have shown that in an asymptotic model for Active Learning where one bounds the number of queries the algorithm makes before it finds a good function (i.e. one of arbitrarily small error rate), but not the number of queries before it can prove or it knows it has found a good function, one can obtain significantly better bounds on the number of label queries required to learn.

$\mathrm{err}_{D,O}(h)$ 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 $X \times Y$, for all $0 < \delta < 1/2$ and $m \in \mathbb{N}$,*

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

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

Classic examples of such subroutines are the (distribution independent) VC bound [35] and the Occam Razor bound [12], or the newer data dependent generalization bounds such as those based on Rademacher Complexities [13]. For concreteness, a VC bound subroutine is stated in Appendix A.

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 [19]. 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, a hypothesis cannot be eliminated based on its disagreement with a single example. Any algorithm must be more conservative without risking eliminating the 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:

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

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

It's important to understand that the ability to sample from the unlabeled data distribution D implies that ability to compute $\text{DISAGREE}_D(H_i)$. To see this, note that: $\text{DISAGREE}_D(H_i) = E_{x\sim D} I(\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x))$ is an expectation over unlabeled points drawn from D. Consequently, Chernoff bounds on the empirical expectation of a $\{0, 1\}$ random variable imply that $\text{DISAGREE}_D(H_i)$ can be estimated to any desired precision with probability 1 using an unlabeled dataset with size limiting to infinity.

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 fresh set of examples S 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 is large enough to eliminate at least half of the current region of uncertainty.

Since A^2 doesn't label 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 $\text{DISAGREE}_D(H_i)$ 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, A^2 uses a schedule of δ according to the following rule: the kth bound evaluation has confidence $\delta_k = \frac{\delta}{k(k+1)}$, for $k \geq 1$. In Algorithm 1, k keeps track of the number of bound computations and i of the number of rounds.

Note: It is important to note that A^2 does not need to know η in advance. Similarly, it does not need to know D in advance.

3.1 Correctness

Theorem 3.1 (Correctness) *For all H, for all (D, O), for all valid subroutines for computing UB and LB, for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, with probability $1 - \delta$, A^2 returns an ϵ-optimal hypothesis or does not terminate.*

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

set $i \leftarrow 1$, $D_i \leftarrow D$, $H_i \leftarrow H$, $H_{i-1} \leftarrow H$, $S_{i-1} \leftarrow \emptyset$, and $k \leftarrow 1$.

(1) **while** $\textsc{Disagree}_D(H_{i-1})\,(\min_{h \in H_{i-1}} \mathrm{UB}(S_{i-1}, h, \delta_k) - \min_{h \in H_{i-1}} \mathrm{LB}(S_{i-1}, h, \delta_k)) > \epsilon$

set $S_i \leftarrow \emptyset$, $H'_i \leftarrow H_i$, $k \leftarrow k+1$

(2) **while** $\textsc{Disagree}_D(H'_i) \geq \frac{1}{2}\textsc{Disagree}_D(H_i)$

if $\textsc{Disagree}_D(H_i)\,(\min_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k) - \min_{h \in H_i} \mathrm{LB}(S_i, h, \delta_k)) \leq \epsilon$

$(*)$ **return** $h = \mathrm{argmin}_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k)$.

else $S'_i =$ rejection sample $2|S_i|+1$ samples x from D satisfying

$$\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x).$$

$S_i \leftarrow S_i \cup \{(x, O(x)) : x \in S'_i\}$, $k \leftarrow k+1$

$(**)$ $H'_i = \{h \in H_i : \mathrm{LB}(S_i, h, \delta_k,) \leq \min_{h' \in H_i} \mathrm{UB}(S_i, h', \delta_k)\}$,

$k \leftarrow k+1$

end if

end while

$H_{i+1} \leftarrow H'_i$, $D_{i+1} \leftarrow D_i$ restricted to $\{x : \exists h_1, h_2 \in H'_i : h_1(x) \neq h_2(x)\}$

$i \leftarrow i+1$

end while

return $h = \mathrm{argmin}_{h \in H_{i-1}} \mathrm{UB}(S_{i-1}, h, \delta_k)$.

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: The first claim is that all bound evaluations are valid simultaneously with probability at least $1-\delta$, and the second is 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

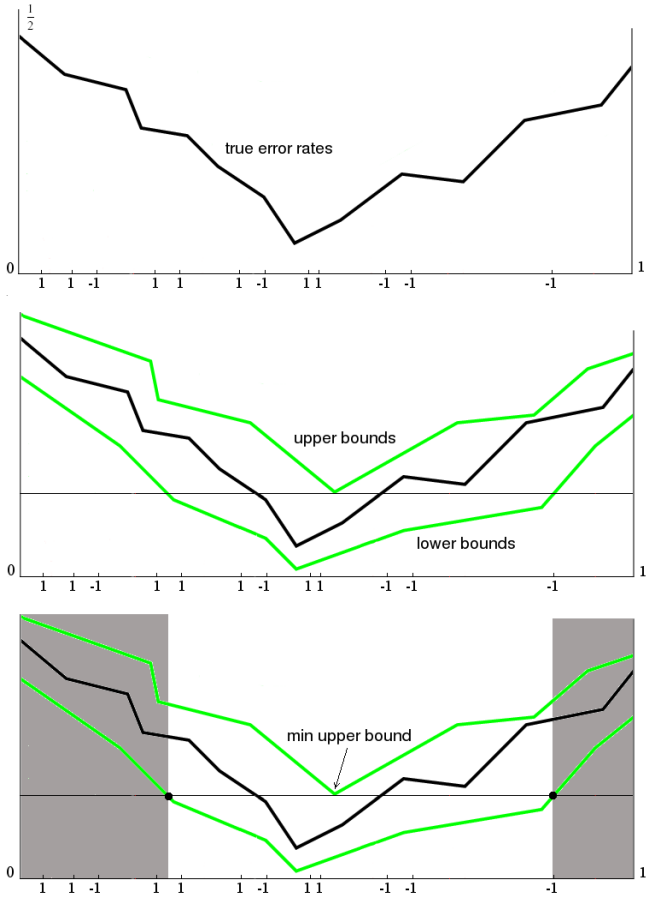

Fig. 3.1. A^2 in action: Sampling, Bounding, Eliminating.

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, step $(**)$ never eliminates a hypothesis that has minimum error rate with respect (D, O).

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 $G \subseteq H$, the error rate of every hypothesis h decomposes 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, finding an ϵ-optimal hypothesis requires only bounding $\mathrm{err}_{D_G,O}(h) \cdot \textsc{Disagree}_D(G)$ to precision ϵ. But this is exactly what the negation of the main while-loop guard does, and this is also the condition used in the first step of the second while loop of the algorithm. In other words, upon termination A^2 satisfies

$$\textsc{Disagree}_D(H_i)(\min_{h \in H_i} \mathrm{UB}(S_i, h, \delta_k) - \min_{h \in H_i} \mathrm{LB}(S_i, h, \delta_k)) \leq \epsilon,$$

which proves the desired result. ∎

3.2 Fall-back Analysis

This section shows that A^2 is never much worse than a standard batch, bound-based algorithm in terms of the number of samples required in order to learn. (A standard example of a bound-based learning algorithm is Empirical Risk Minimization (ERM) [36].)

The sample complexity $m(\epsilon, \delta, H)$ required by a batch algorithm that uses a subroutine for computing $\mathrm{LB}(S, h, \delta)$ and $\mathrm{UB}(S, h, \delta)$ is defined as the minimum number of samples m such that for all $S \in X^m$, $|\mathrm{UB}(S, h, \delta) - \mathrm{LB}(S, h, \delta)| \leq \epsilon$ for all $h \in H$. For concreteness, this section uses the following bound on $m(\epsilon, \delta, H)$ stated as Theorem A.1 in Appendix A:

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

Here V_H 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 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 [35] and the Occam's Razor bound [12].

Theorem 3.2 *For all H, for all (D, O), for all UB and LB satisfying the assumption above, for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, 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. The proof works by finding an upper bound $N(\epsilon, \delta, H)$ on the number of bound evaluations throughout the life of the algorithm. This implies that the confidence parameter δ_k is always greater than $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$.

Recall that D_i is the distribution over x used on the ith iteration of the first while loop.

Consider $i = 1$. If condition 2 of Algorithm A^2 is repeatedly satisfied then after labeling $m(\epsilon, \delta', H)$ examples from D_1 for all hypotheses $h \in H_1$,

$$|\mathrm{UB}(S_1, h, \delta') - \mathrm{LB}(S_1, h, \delta')| \leq \epsilon$$

simultaneously. Note that in these conditions A^2 safely halts. Notice also that the number of bound evaluations during this process is at most $\log_2 m(\epsilon, \delta', H)$.

On the other hand, if loop (2) ever completes and i increases, then it is enough, if you finish when $i = 2$, 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 in the outer while-loop and the 'if' in Step 2 of A^2.) Uniformly bounding the gap between upper and lower bounds over all hypotheses $h \in H_2$ to within 2ϵ, requires $m(2\epsilon, \delta', H) \leq \frac{m(\epsilon, \delta', H)}{2}$ labeled examples from D_2 and the number of bound evaluations in round $i = 2$ is at most $\log_2 m(\epsilon, \delta', H)$.

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 which requires $m(2^{i-1}\epsilon, \delta', H) \leq \frac{m(\epsilon, \delta', H)}{2^{i-1}}$ labeled examples from D_i. Also the number of bound evaluations in round i is at most $\log_2 m(\epsilon, \delta', H)$.

Since the number of rounds is bounded by $\log_2 \frac{1}{\epsilon}$, it follows that the maximum number of bound evaluations throughout the life of the algorithm is at most $\log_2 \frac{1}{\epsilon} \log_2 m(\epsilon, \delta', H)$. This implies that in order to determine an upper bound $N(\epsilon, \delta, H)$ only a solution to the inequality:

$$N(\epsilon, \delta, H) \geq \log_2 \frac{1}{\epsilon} \log_2 m\left(\epsilon, \frac{\delta}{N(\epsilon, \delta, H)(N(\epsilon, \delta, H) + 1)}, H\right)$$

is required.

Finally, adding up the number of calls to the label oracle O in all rounds yields at most $2m(\epsilon, \delta', H)$ over the life of the algorithm. ∎

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

Corollary 3.3 *For all hypothesis classes H of VC-dimension V_H, for all distributions (D, O) over $X \times Y$, for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, the algorithm A^2 requires at most $\tilde{O}\left(\frac{1}{\epsilon^2}(V_H \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\right)$ labeled examples the oracle O.* [2]

Proof: The form of $m(\epsilon, \delta, H)$ and Theorem 3.2 implies an upper bound on $N = N(\epsilon, \delta, H)$. It is enough to find the smallest N satisfying

$$N \geq \ln\left(\frac{1}{\epsilon}\right) \ln\left(\frac{64}{\epsilon^2}\left(2V_H \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4N^2}{\delta}\right)\right)\right).$$

Using the inequality $\ln a \leq ab - \ln b - 1$ for all $a, b > 0$ and some simple algebraic manipulations, the desired upper bound on $N(\epsilon, \delta, H)$ holds. The result then follows from Theorem 3.2. ∎

4 Active Learning Speedups

This section gives examples of exponential sample complexity improvements achieved by A^2.

4.1 Learning Threshold Functions

Linear threshold functions are the simplest and easiest to analyze class. It turns out that even for this class, exponential reductions in sample complexity are not achievable when the noise rate η is large [29]. We prove the following three results:

(1) An exponential improvement in sample complexity when the noise rate is small (Theorem 4.1).
(2) A slower improvement when the noise rate is large (Theorem 4.2). In the extreme case where the noise rate is 1/2, there is no improvement.
(3) An exponential improvement when the noise rate is large but due to constant label noise (Theorem 4.3). This shows that for some forms of high noise exponential improvement remains possible.

All results in this section assume that subroutines LB and UB in A^2 are based on the VC bound.

[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.

Theorem 4.1 *Let H be the set of thresholds on an interval. For all distributions (D, O) where D is a continuous probability distribution function, for any $\epsilon < \frac{1}{2}$ and $\frac{\epsilon}{16} \geq \eta$, 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 $1-\delta$.*

Proof: Consider round $i \geq 1$ of the algorithm. 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 the distribution D_i, i.e., $d_i(h_1, h_2) = \mathbf{Pr}_{x\sim D_i}[h_1(x) \neq h_2(x)]$.

Let h^* be any minimum error rate hypothesis in H. Note that for any hypothesis $h \in H_i$, we have $\mathrm{err}_{D_i,O}(h) \geq d_i(h, h^*) - \mathrm{err}_{D_i,O}(h^*)$ and $\mathrm{err}_{D_i,O}(h^*) \leq \eta/Z_i$, where $Z_i = \mathbf{Pr}_{x\sim D}[x \in [lower_i, upper_i]\,]$ is a shorthand for $\textsc{Disagree}_D(H_i)$ and $[lower_i, upper_i]$ denotes the support of D_i. Thus $\mathrm{err}_{D_i,O}(h^*) \leq d_i(h, h^*) - \eta/Z_i$.

We will show 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 $[lower_i, upper_i]$. 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 $d_i(h^*, u_i) = \frac{1}{4}$. 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 suppose 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 all $h \in [u_i, upper_i]$ satisfy $d_i(h^*, h) \geq \frac{1}{4}$ and moreover $\mathbf{Pr}_{x\sim D_i}[x \in [u_i, upper_i]\,] \geq \frac{1}{2}$. A similar argument holds for $d_i(h^*, upper_i) \leq \frac{1}{4}$.

Using the VC bound, 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 for all hypotheses $h \in H_i$ simultaneously, $|\mathrm{UB}(S_i, h, \delta) - \mathrm{LB}(S_i, h, \delta)| \leq \frac{1}{8} - \frac{\eta}{Z_i}$ holds. Note that η/Z_i is always upper bounded by $\frac{1}{16}$.

Consider a hypothesis $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. For any such h, $\mathrm{err}_{D_i,O}(h) \geq d_i(h, h^*) - \eta/Z_i \geq \frac{1}{4} - \frac{\eta}{Z_i}$, 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, $\text{err}_{D_i,O}(h^*) \leq \frac{\eta}{Z_i}$, and so $\text{UB}(S_i, h^*, \delta) \leq \frac{\eta}{Z_i} + \frac{1}{8} - \frac{\eta}{Z_i} = \frac{1}{8}$. Thus A^2 eliminates all $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. But that means $\textsc{Disagree}_D(H_i') \leq \frac{1}{2}\textsc{Disagree}_D(H_i)$, thus terminating round i.[3]

Each exit from **while** loop (2) decreases $\textsc{Disagree}_D(H_i)$ by at least a factor of 2, implying that the number of executions is bounded by $\log \frac{1}{\epsilon}$. The algorithm makes $O\left(\ln\left(\frac{1}{\delta'}\right)\ln\left(\frac{1}{\epsilon}\right)\right)$ calls to the oracle, where $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and $N(\epsilon, \delta, H)$ is an upper bound on the number of bound evaluations throughout the life of the algorithm.

The number of bound evaluations required in round i is $O\left(\ln \frac{1}{\delta'}\right)$, which implies that $N(\epsilon, \delta, H)$ should satisfy $c \ln\left(\frac{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}{\delta}\right)\ln\left(\frac{1}{\epsilon}\right) \leq N(\epsilon, \delta, H)$, for some constant c. Solving this inequality completes the proof. ∎

Theorem 4.2 below asymptotically matches a lower bound of Kääriäinen [29]. Recall that A^2 does not need to know η in advance.

Theorem 4.2 *Let H be the set of thresholds on an interval. Suppose that $\epsilon < \frac{1}{2}$ and $\eta > 16\epsilon$. For all D, with probability $1-\delta$, the algorithm A^2 requires at most $\tilde{O}\left(\frac{\eta^2 \ln \frac{1}{\delta}}{\epsilon^2}\right)$ labeled samples.*

Proof: The proof is similar to the previous proof. Theorem 4.1 implies that loop (2) completes $\Theta(\log \frac{1}{\eta})$ times. At this point, the minimum error rate of the remaining hypotheses conditioned on disagreement becomes sufficient so that the algorithm may only halt via the return step $(*)$. In this case, $\textsc{Disagree}_D(H) = \Theta(\eta)$ implying that the number of samples required is $\tilde{O}\left(\frac{\eta^2 \ln \frac{1}{\delta}}{\epsilon^2}\right)$. ∎

The final theorem is for the constant noise case where $|\mathbf{Pr}_{y\sim O|x}[h^*(x) \neq y] - \frac{1}{2}| = \eta$ for all $x \in X$. The theorem is similar to earlier work [15], except that we achieve these improvements with a general purpose active learning algorithm that does not use any prior over the hypothesis space or knowledge of the noise rate, and is applicable to arbitrary hypothesis spaces.

Theorem 4.3 *Let H be the set of thresholds on an interval. For all unlabeled data distributions D, for all labeled data distributions O, for any constant label noise $\eta < 1/2$ and $\epsilon < \frac{1}{2}$, the algorithm A^2 makes $O\left(\frac{1}{(1-2\eta)^2}\ln\left(\frac{1}{\epsilon}\right)\ln\left(\frac{\ln\left(\frac{1}{\epsilon\delta}\right)}{\delta}\right)\right)$ calls to the oracle O on examples drawn* i.i.d. *from D, with probability $1-\delta$.*

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

The proof is essentially the same as for Theorem 4.1, except that the constant label noise condition implies that the amount of noise in the remaining actively labeled subset stays bounded through the recursions.

Proof: Consider round $i \geq 1$. For $h_1, h_2 \in H_i$, let $d_i(h_1, h_2) = \mathbf{Pr}_{x \sim D_i}[h_1(x) \neq h_2(x)]$. Note that for any hypothesis $h \in H_i$, we have $\mathrm{err}_{D_i,O}(h) = d_i(h, h^*)(1-2\eta) + \eta$ and $\mathrm{err}_{D_i,O}(h^*) = \eta$, where h^* is a minimum error rate threshold.

As in the proof of Theorem 4.1, 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 support $[lower_i, upper_i]$ of D_i.

The VC bound implies that for any $\delta' > 0$ with probability $1 - \delta'$, if $|S_i| = O\left(\frac{\ln(1/\delta')}{(1-2\eta)^2}\right)$, then for all hypotheses $h \in H_i$ simultaneously, $|\mathrm{UB}(S_i, h, \delta) - \mathrm{LB}(S_i, h, \delta)| < \frac{1-2\eta}{8}$.

Consider a hypothesis $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. For any such h, $\mathrm{err}_{D_i,O}(h) \geq \frac{1-2\eta}{4} + \eta = \frac{1}{4} + \frac{\eta}{2}$, and so $\mathrm{LB}(S_i, h, \delta) > \frac{1}{4} + \frac{\eta}{2} - \frac{1}{8}(1-2\eta) = \frac{1}{8} + \frac{3\eta}{4}$. On the other hand, $\mathrm{err}_{D_i,O}(h^*) = \eta$, and so $\mathrm{UB}(S_i, h^*, \delta) < \eta + (\frac{1}{8} - \frac{\eta}{4}) = \frac{1}{8} + \frac{3\eta}{4}$. Thus A^2 eliminates all $h \in H_i$ with $d_i(h, h^*) \geq \frac{1}{4}$. But this means that $\textsc{Disagree}_D(H_i') \leq \frac{1}{2}\textsc{Disagree}_D(H_i)$, thus terminating round i.

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 $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and $N(\epsilon, \delta, H)$ is an upper bound on the number of bound evaluations throughout the life of the algorithm. The number of bound evaluations required in round i is $O(\ln(1/\delta'))$, which implies that the number of bound evaluations throughout the life of the algorithm $N(\epsilon, \delta, H)$ should satisfy $c \ln\left(\frac{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}{\delta}\right)\ln\left(\frac{1}{\epsilon}\right) \leq N(\epsilon, \delta, H)$, for some constant c. Solving this inequality, completes the proof. ∎

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. Note that even in this seemingly simple scenario, there exists an $\Omega\left(\frac{1}{\epsilon}\left(d + \log\frac{1}{\delta}\right)\right)$ lower bound on the PAC passive supervised learning sample complexity [31]. We will show that A^2 provides exponential savings in this case even in the presence of arbitrary forms 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 $h(x) = \text{sign}(w \cdot x)$. 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 $d_D(u,v) = \frac{\arccos(u \cdot v)}{\pi}$. Finally, let $\theta(u,v) = \arccos(u \cdot v)$. Thus $d_D(u,v) = \frac{\theta(u,v)}{\pi}$.

Theorem 4.4 *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 $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and

$$N(\epsilon,\delta,H) = O\left(\ln\frac{1}{\epsilon}\left(d^2\ln d + d\ln\frac{d\ln\frac{1}{\epsilon}}{\delta}\right)\right).$$

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 $\mathbf{Pr}_{x\sim D}[x \in R_i] = \textsc{Disagree}_D(H_i)$.

Consider round i of A^2. We prove that the round completes with high probability if a certain threshold on the number of labeled examples is reached. The round may complete with a smaller number of examples, but this is fine because the metric of progress $\textsc{Disagree}_D(H_i)$ must halve in order to complete.

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 ith distribution D_i to guarantee, with probability $1-\delta'$, that for all $w \in H_i$,

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

where r_i is a shorthand for $\textsc{Disagree}_D(H_i)$. (By assumption, $\eta < \frac{\epsilon}{16\sqrt{d}}$ and the loop guard guarantees that $\textsc{Disagree}_D(H_i) \geq \epsilon$. Thus the precision above is at least $\frac{1}{32\sqrt{d}}$.)[4] This implies that $\text{UB}(S,w,\delta') - \operatorname{err}_{D_i,O}(w) < \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i}$, and $\operatorname{err}_{D_i,O}(w) - \text{LB}(S,w,\delta') < \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i}$. Consider any $w \in H_i$ with $d_{D_i}(w,w^*) \geq \frac{1}{4\sqrt{d}}$. For any such w, $\operatorname{err}_{D_i,O}(w) \geq \frac{1}{4\sqrt{d}} - \frac{\eta}{r_i}$, and so

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

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

However, $\mathrm{err}_{D_i,O}(w^*) \le \frac{\eta}{r_i}$, and thus $\mathrm{UB}(S, w^*, \delta') < \frac{\eta}{r_i} + \frac{1}{8\sqrt{d}} - \frac{\eta}{r_i} = \frac{1}{8\sqrt{d}}$, so A^2 eliminates w in step $(**)$.

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

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

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

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 $\cos\theta(w^*, x) \le \cos(\frac{\pi}{2} - \theta(w, w^*)) = \sin\theta(w, w^*) < \frac{r_i}{2\sqrt{d}}$, where the last inequality is due to the fact that A^2 eliminates all w with $\sin\theta(w, w^*) \ge \frac{r_i}{2\sqrt{d}}$. Thus any $x \in R_{i+1}$ must satisfy $|w^* \cdot x| < \frac{r_i}{2\sqrt{d}}$.

Using the fact that $\mathbf{Pr}[A \mid B] = \frac{\mathbf{Pr}[AB]}{\mathbf{Pr}[B]} \le \frac{\mathbf{Pr}[A]}{\mathbf{Pr}[B]}$ for any A and B,

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

where the third inequality follows from Lemma A.2. Thus $\textsc{Disagree}_D(H_{i+1}) \le \frac{1}{2}\textsc{Disagree}_D(H_i)$, as desired.

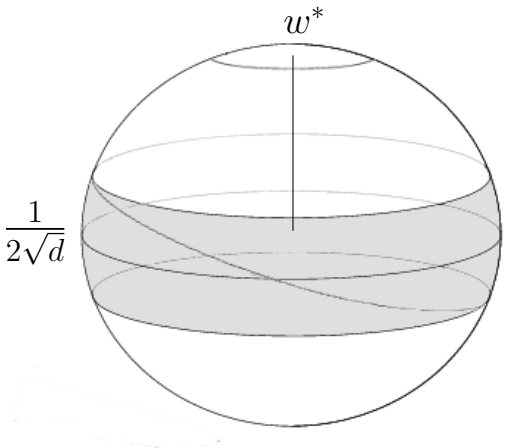

Fig. 4.1. The region of uncertainty after the first iteration (schematic).

In order to finish the argument, it suffices to notice that since every round cuts $\textsc{Disagree}_D(H_i)$ 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 $\delta' = \frac{\delta}{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}$ and $N(\epsilon,\delta,H)$ is an upper bound on the number of bound evaluations throughout the life of the algorithm. 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 $c\left(d^2 \ln d + d \ln\left(\frac{N(\epsilon,\delta,H)(N(\epsilon,\delta,H)+1)}{\delta}\right)\right) \ln\left(\frac{1}{\epsilon}\right) \leq N(\epsilon,\delta,H)$ for some constant c. Solving this inequality, completes the proof. ∎

Note: For comparison, the query complexity of the Perceptron-based active learner of [24] 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 $\eta = 0$. Similar bounds are obtained in [5] both in the realizable case and for a specific form of noise related to the Tsybakov small noise condition. The cleanest and simplest argument that exponential improvement is in principle possible in the realizable case for the same H, X, and D appears in [22]. Our work provides the first justification of why one can hope to achieve similarly strong guarantees in the much harder agnostic case, when the noise rate is sufficiently small with respect to the desired error.

5 Conclusions, Discussion and Open Questions

This paper presents A^2, the first active learning algorithm that finds an ϵ-optimal hypothesis in any hypothesis class, when the distribution has arbitrary forms of noise. The algorithm relies only upon the assumption that the samples are drawn *i.i.d.* from a fixed (unknown) distribution, and it does not need to know the error rate of the best classifier in the class in advance. We analyze A^2 for several settings considered before in the realizable case, showing that A^2 achieves an exponential improvement over the usual sample complexity of supervised learning in these settings. We also provide a guarantee that A^2 never requires substantially more labeled examples than passive learning.

Subsequent Work Following the initial publication of A^2, Hanneke has further analyzed the A^2 algorithm [26], deriving a general upper bound on the number of label requests made by A^2. This bound is expressed in terms of particular quantity called the *disagreement coefficient*, which roughly quantifies how quickly the region of disagreement can grow as a function of the radius of the version space. For concreteness the bound is included in Appendix B.

In addition, Dasgupta, Hsu, and Monteleoni [23] introduce and analyze a new agnostic active learning algorithm. While similar to A^2, this algorithm simplifies the maintenance of the region of uncertainty with a reduction to supervised learning, keeping track of the version space implicitly via label constraints.

Open Questions A common feature of the selective sampling algorithm [19], A^2, and others [23] is that they are all non-aggressive in their choice of query

points. Even points on which there is a small amount of uncertainty are queried, rather than pursuing the maximally uncertain point. In recent work Balcan, Broder and Zhang [5] have shown that more aggressive strategies can generally lead to better bounds. However the analysis in [5] was specific to the realizable case, or done for a very specific type of noise. It is an open question to design aggressive agnostic active learning algorithms.

A more general open question is what conditions are sufficient and necessary for active learning to succeed in the agnostic case. What is the right quantity that can characterize the sample complexity of agnostic active learning? As mentioned already, some progress in this direction has been recently made in [26] and [23]; however, those results characterize non-aggressive agnostic active learning. Deriving and analyzing the optimal agnostic active learning strategy is still an open question.

Much of the existing literature on active learning has been focused on binary classification; it would be interesting to analyze active learning for other loss functions. The key ingredient allowing recursion in the proof of correctness is a loss that is unvarying with respect to substantial variation over the hypothesis space. Many losses such as squared error loss do not have this property, so achieving substantial speedups, if that is possible, requires new insights. For other losses with this property (such as hinge loss or clipped squared loss), generalizations of A^2 appear straightforward.

Acknowledgements Maria-Florina has been supported in part by the NSF grant CCF-0514922, by an IBM Graduate Fellowship, and by a Google Research Grant. Part of this work was done while she was visiting IBM T.J. Watson Research Center in Hawthorne, NY, and Toyota Technological Institute at Chicago.

References

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1998.

[2] D. Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

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

[4] M.-F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

[5] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In

Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT), 2007.

[6] M.-F. Balcan, E. Even-Dar, S. Hanneke, M. Kearns, Y. Mansour, and J. Wortman. Asymptotic active learning. In *Workshop on Principles of Learning Design Problem*. in conjunction with the 21st Annual Conference on Neural Information Processing Systems (NIPS), 2007.

[7] M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*, 2008.

[8] E. Baum and K. Lang. Query learning can work poorly when a human oracle is used. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1993.

[9] A. Blum. Machine learning theory. Essay, 2007.

[10] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 19–26, 2001.

[11] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *The 11th Annual Conference on Computational Learning Theory (COLT)*, 1998.

[12] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

[13] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of Classification: A Survey of Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[14] N. H. Bshouty and N. Eiron. Learning monotone dnf from a teacher that almost does not answer membership queries. *Journal of Machine Learning Research*, 3:49–57, 2002.

[15] M. Burnashev and K. Zigangirov. An interval estimation problem for controlled observations. *Problems in Information Transmission*, 10:223–231, 1974.

[16] R. Castro and R. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

[17] R. Castro, R. Willett, and R. Nowak. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[18] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. MIT press, 2006.

[19] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 201–221, 1994.

[20] J. Czyzowicz, D. Mundici, and A. Pelc. Ulam's searching game with lies. *Journal of Combinatorial Theory*, 52:62–76, 1989.

[21] S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems*, 17, 2004.

[22] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[23] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007.

[24] S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

[25] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

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

[27] J. Jackson. An efficient membership-query algorithm for learning dnf with respect to the uniform distribution. *Journal of Computer and System Sciences*, 57(3):414–440, 1995.

[28] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 200–209, 1999.

[29] M. Kääriäinen. On active learning in the non-realizable case. In *Proceedings of 17th International Conference on Algorithmic Learning Theory (ALT)*, volume 4264 of *Lecture Notes in Computer Science*, pages 63–77, 2006.

[30] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *Proceedings of the 9th International Conference on Computer Vision (ICCV)*, pages 626–633, 2003.

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

[32] T. Mitchell. The discipline of machine learning. Technical Report CMU ML-06 108, 2006.

[33] K. Nigam, A. McCallum, S. Thrun, and T.M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[34] S. B. Park and B. T. Zhang. Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. *Information Processing and Management*, 40(3):421–439, 2004.

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

[36] V. N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, 1998.

A Standard Results

The following standard Sample Complexity bound is in [3].

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

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

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

Section 4.2 uses the following a classic lemma about the uniform distribution. For a proof see, for example, [5,24].

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.

B Subsequent Guarantees for A^2

This section describes the disagreement coefficient [26] and the guarantees it provides for the A^2 algorithm. We begin with a few additional definitions, in the notation of Section 2.

Definition 2 *The* disagreement rate $\Delta(V)$ *of a set $V \subseteq H$ is defined as*

$$\Delta(V) = \mathbf{Pr}_{x \sim D}[x \in \textsc{Disagree}_D(V)].$$

Definition 3 *For $h \in H$, $r > 0$, let $B(h, r) = \{h' \in H : d(h', h) \leq r\}$ and define the* disagreement rate at radius r *as*

$$\Delta_r = \sup_{h \in H}\left(\Delta(B(h, r))\right).$$

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

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

Theorem B.1 *[26] If* θ *is the disagreement coefficient for* H*, then with probability at least* $1 - \delta$*, given the inputs* ϵ *and* δ*,* A^2 *outputs an* ϵ*-optimal hypothesis* h*. Moreover, the number of label requests made by* A^2 *is at most:*

$$\tilde{O}\left(\theta^2 \left(\frac{\eta^2}{\epsilon^2} + 1\right)\left(V_H \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}\right) \ln \frac{1}{\epsilon}\right),$$

where $V_H \geq 1$ *is the VC-dimension of* H*.*