A general agnostic active learning algorithm

Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni
University of California, San Diego
{dasgupta,djhsu,cmontel}@cs.ucsd.edu

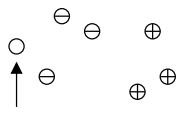
Abstract

We present a simple, agnostic active learning algorithm that works for any hypothesis class of bounded VC dimension, and any data distribution. Our algorithm extends a scheme of Cohn, Atlas, and Ladner to the agnostic setting, by (1) reformulating it using a reduction to supervised learning and (2) showing how to apply generalization bounds even for the non-i.i.d. samples that result from selective sampling. We provide a general characterization of the label complexity of our algorithm. This quantity is never more than the usual PAC sample complexity of supervised learning, and is exponentially smaller for some hypothesis classes and distributions. We also demonstrate improvements experimentally.

Introduction

Active learning addresses the issue that, in many applications, labeled data typically comes at a higher cost (e.g. in time, effort) than unlabeled data. An active learner is given unlabeled data and must pay to view any label. The hope is that significantly fewer labeled examples are used than in the supervised (non-active) learning model. Active learning applies to a range of data-rich problems such as genomic sequence annotation and speech recognition. In this paper we formalize, extend, and provide label complexity guarantees for one of the earliest and simplest approaches to active learning—one due to Cohn, Atlas, and Ladner (1994).

The scheme of Cohn, Atlas, and Ladner examines data one by one in a stream and requests the label of any data point about which it is currently unsure. For example, suppose the hypothesis class consists of linear separators in the plane, and assume that the data is linearly separable. Let the first six data be labeled as follows.



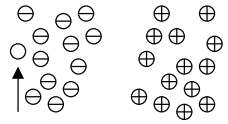
The learner does not need to request the label of the seventh point (indicated by the arrow) because it is not unsure about the label: any straight line with the $\oplus$s and $\ominus$s on opposite sides has the seventh point with the $\ominus$s. Put another way, the point is not in the *region of uncertainty* (Cohn, Atlas, and Ladner, 1994), the portion of the data space for which there is disagreement among hypotheses consistent with the present labeled data.

Although very elegant and intuitive, this approach to active learning faces two problems:

1. Explicitly maintaining the region of uncertainty can be computationally cumbersome.
2. Data is usually not perfectly separable.

Our main contribution is to address these problems. We provide a simple generalization of the selective sampling scheme of Cohn, Atlas, and Ladner that tolerates adversarial noise and never requests many more labels than a standard agnostic supervised learner would to learn a hypothesis with the same error.

In the previous example, an *agnostic* active learner (one that does not assume a perfect separator exists) is actually *still* uncertain about the label of the seventh point, because all six of the previous labels could be inconsistent with the best separator. Therefore, it should still request the label. On the other hand, after enough points have been labeled, if an unlabeled point occurs at the position shown below, chances are its label is not needed.



To extend the notion of uncertainty to the agnostic setting, we divide the observed data points into two groups, $\hat{S}$ and T:

- Set $\hat{S}$ contains the data for which we did *not* request labels. We keep these points around and assign them the label we think they should have.
- Set T contains the data for which we explicitly requested labels.

We will manage things in such a way that the data in $\hat{S}$ are always consistent with the best separator in the class. Thus, somewhat counter-intuitively, the labels in $\hat{S}$ are completely reliable whereas the labels in T could be inconsistent with the best separator. To decide whether we are uncertain about the label of a new point x, we reduce to supervised learning: we learn hypotheses h_{+1} and h_{-1} such that

- h_{+1} is consistent with all the labels in $\hat{S} \cup \{(x, +1)\}$ and has minimal empirical error on T, while
- h_{-1} is consistent with all the labels in $\hat{S} \cup \{(x, -1)\}$ and has minimal empirical error on T.

If, say, the *true* error of the hypothesis h_{+1} is much larger than that of h_{-1}, we can safely infer that the best separator must also label x with -1 without requesting a label; if the error difference is only modest, we explicitly request a label. Standard generalization bounds for an i.i.d. sample let us perform this test by comparing *empirical* errors on $\hat{S} \cup T$.

The last claim may sound awfully suspicious, because $\hat{S} \cup T$ is not i.i.d.! Indeed, this is in a sense the core sampling problem that has always plagued active learning: the labeled sample T might not be i.i.d. (due to the filtering of examples based on an adaptive criterion), while $\hat{S}$ only contains unlabeled examples (with made-up labels). Nevertheless, we prove that in our case, it is in fact correct to effectively pretend $\hat{S} \cup T$ is an i.i.d. sample. A direct consequence is that the *label complexity* of our algorithm (the number of labels requested before achieving a desired error) is never much more than the usual sample complexity of supervised learning (and in some cases, is significantly less).

An important algorithmic detail is the specific choice of generalization bound we use in deciding whether to request a label or not. A small polynomial difference in generalization rates (between $n^{-1/2}$ and n^{-1}, say) can get magnified into an *exponential* difference in label complexity, so it is crucial for us to use a good bound. We use a normalized bound that takes into account the empirical error (computed on $\hat{S} \cup T$—again, not an i.i.d. sample) of the hypothesis in question.

Earlier work on agnostic active learning (Balcan, Beygelzimer, and Langford, 2006; Hanneke, 2007) has been able to upper bound label complexity in terms of a parameter of the hypothesis class (and data distribution) called the *disagreement coefficient*. We give label complexity bounds for our method based on this same quantity, and we get a better dependence on it, linear rather than quadratic.

To summarize, in this paper we present and analyze a simple agnostic active learning algorithm for general hypothesis classes of bounded VC dimension. It extends the selective sampling scheme of Cohn, Atlas, and Ladner (1994) to the agnostic setting, using normalized generalization bounds, which we apply in a simple but subtle manner. For certain hypothesis classes and distributions, our analysis yields improved label complexity guarantees over the standard sample complexity of supervised learning. We also demonstrate such improvements experimentally.

Related work

A large number of algorithms have been proposed for active learning, under a variety of learning models. In this section, we consider only methods whose generalization behavior has been rigorously analyzed.

An early landmark result was the selective sampling scheme of Cohn, Atlas, and Ladner (1994) described above. This simple active learning algorithm, designed for separable data, has been the inspiration for a lot of subsequent work. A few years later, the seminal work of Freund, Seung, Shamir, and Tishby (1997) analyzed an algorithm called *query-by-committee* that operates in a Bayesian setting and uses an elegant sampling trick for deciding when to query points. The core primitive required by this algorithm is the ability to sample randomly from the posterior over the hypothesis space. In some cases this can be achieved efficiently (Gilad-Bachrach, Navot, and Tishby, 2005), for instance when the hypothesis class consists of linear separators in $\mathbb{R}^d$ (with a uniform prior) and the data is distributed uniformly over the surface of the unit sphere in $\mathbb{R}^d$. In this particular setting, the authors showed that the number of labels required to achieve generalization error ε is just $O(d \log 1/\varepsilon)$, exponentially lower than the usual supervised sample complexity of $O(d/\varepsilon)$.

Subsequently, Dasgupta, Kalai, and Monteleoni (2005) showed that a simple variant of the perceptron algorithm also achieves this label complexity, even for a worst-case (non-Bayesian) choice of target hypothesis.

All the work mentioned so far assumes *separable* data. This case was studied abstractly by Dasgupta (2005), who found that a parameter called the *splitting index* loosely characterizes the label complexity of actively learning hypothesis classes of bounded VC dimension. As yet, it is not known how to realize this label complexity in a computationally efficient way, except in special cases.

A natural way to formulate active learning in the *agnostic* setting is to ask the learner to return a hypothesis with error at most $\nu + \varepsilon$ (where ν is the error of the best hypothesis in the specified class) using as few labels as possible. A basic constraint on the label complexity was pointed out by Kääriäinen (2006), who showed that for any $\nu \in (0, 1/2)$, there are data distributions that force any active learner that achieves error at most $\nu + \varepsilon$ to request $\Omega((\nu/\varepsilon)^2)$ labels.

The first rigorously-analyzed agnostic active learning algorithm, called A^2, was developed recently by Balcan, Beygelzimer, and Langford (2006). Like Cohn-Atlas-Ladner (1994), this algorithm uses a region of uncertainty, although the lack of separability complicates matters and A^2 ends up explicitly maintaining an ε-net of the hypothesis space. Subsequently, Hanneke (2007) characterized the label complexity of the A^2 algorithm in terms of a parameter called the *disagreement coefficient*. Another thread of work focuses on agnostic learning of thresholds for data that

lie on a line; in this case, a precise characterization of label complexity can be given (Castro and Nowak, 2006, 2007).

These previous results either make strong distributional assumptions (such as separability, or a uniform input distribution), or else they are computationally prohibitive in general.

Our work was inspired by both Cohn-Atlas-Ladner and Balcan-Beygelzimer-Langford, and we have built heavily upon their insights. We bound the label complexity of our method in terms of the same parameter as used for A^2 (Hanneke, 2007), and get a somewhat better dependence (linear rather than quadratic).

A common feature of Cohn-Atlas-Ladner, A^2, and our method is that they are all fairly non-aggressive in their choice of query points. They are content with querying all points on which there is even a small amount of uncertainty, rather than, for instance, pursuing the maximally uncertain point. Recently, Balcan, Broder, and Zhang (2007) showed that for the hypothesis class of linear separators, under distributional assumptions on the data (for instance, a uniform distribution over the unit sphere), a more aggressive strategy can yield better label complexity.

Preliminaries

Learning framework and uniform convergence

Let $\mathcal{X}$ be the input space, $\mathcal{D}$ a distribution over $\mathcal{X} \times \{\pm 1\}$ and $\mathcal{H}$ a class of hypotheses $h : \mathcal{X} \to \{\pm 1\}$ with VC dimension $\mathrm{vcdim}(\mathcal{H}) = d < \infty$. Recall that the nth shattering coefficient $\mathcal{S}(\mathcal{H}, n)$ is defined as the maximum number of ways in which $\mathcal{H}$ can label a set of n points; by Sauer's lemma, this is at most $O(n^d)$ (Bousquet, Boucheron, and Lugosi, 2004, p.175). We denote by $\mathcal{D}_{\mathcal{X}}$ the marginal of $\mathcal{D}$ over $\mathcal{X}$. In our active learning model, the learner receives unlabeled data sampled from $\mathcal{D}_{\mathcal{X}}$; for any sampled point x, it can optionally request the label y sampled from the conditional distribution at x. This process can be viewed as sampling (x, y) from $\mathcal{D}$ and revealing only x to the learner, keeping the label y hidden unless the learner explicitly requests it. The error of a hypothesis h under $\mathcal{D}$ is $\mathrm{err}_{\mathcal{D}}(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$, and on a finite sample $Z \subset \mathcal{X} \times \{\pm 1\}$, the empirical error of h is

$$\mathrm{err}(h, Z) \;=\; \frac{1}{|Z|} \sum_{(x,y)\in Z} \mathbb{1}[h(x) \neq y],$$

where $\mathbb{1}[\cdot]$ is the 0-1 indicator function. We assume for simplicity that the minimal error $\nu = \inf\{\mathrm{err}_{\mathcal{D}}(h) : h \in \mathcal{H}\}$ is achieved by a hypothesis $h^* \in \mathcal{H}$.

Our algorithm and analysis use the following normalized uniform convergence bound (Bousquet, Boucheron, and Lugosi, 2004, p.200).

Lemma 1 (Vapnik and Chervonenkis (1971))**.** *Let $\mathcal{F}$ be a family of measurable functions $f : \mathcal{Z} \to \{0, 1\}$ over a space $\mathcal{Z}$. Denote by $\mathbb{E}_Z f$ the empirical average of f over a subset $Z \subset \mathcal{Z}$. Let $\alpha_n = \sqrt{(4/n)\ln(8\mathcal{S}(\mathcal{F}, 2n)/\delta)}$. If Z is an i.i.d. sample of size n from a fixed distribution over $\mathcal{Z}$, then, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$:*

$$\begin{aligned} -\min\left(\alpha_n\sqrt{\mathbb{E}_Z f}, \alpha_n^2 + \alpha_n\sqrt{\mathbb{E} f}\right) \\ \leq\; \mathbb{E} f - \mathbb{E}_Z f \\ \leq\; \min\left(\alpha_n^2 + \alpha_n\sqrt{\mathbb{E}_Z f}, \alpha_n\sqrt{\mathbb{E} f}\right). \end{aligned}$$

Disagreement coefficient

The active learning algorithm we will shortly describe is not very aggressive: rather than seeking out points that are maximally informative, it queries every point that it is somewhat unsure about. The early work of Cohn-Atlas-Ladner (1994) and the recent A^2 algorithm (Balcan, Beygelzimer, and Langford, 2006) are similarly mellow in their querying strategy. The label complexity improvements achievable by such algorithms are nicely captured by a parameter called the *disagreement coefficient*, introduced recently by Hanneke (2007) in his analysis of A^2.

To motivate the disagreement coefficient, imagine that we are in the midst of learning, and that our current hypothesis h_t has error at most β. Suppose we even know the value of β. Then the only candidate hypotheses we still need to consider are those that differ from h_t on at most a 2β fraction of the input distribution, because all other hypotheses must have error more than β. To make this a bit more formal, we impose a (pseudo-)metric ρ on the space of hypotheses, as follows.

Definition 1. *The* disagreement pseudo-metric ρ *on $\mathcal{H}$ is defined by*

$$\rho(h, h') = \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[h(x) \neq h'(x)]$$

for $h, h' \in \mathcal{H}$. Let $B(h, r) = \{h' \in \mathcal{H} : \rho(h, h') \leq r\}$ be the ball centered around h of radius r.

Returning to our earlier scenario, we need only consider hypotheses in $B(h_t, 2\beta)$ and thus, when we see a new data point x, there is no sense in asking for its label if all of $B(h_t, 2\beta)$ agrees on what this label should be. The only points we potentially need to query are

$$\{x : h(x) \neq h'(x) \text{ for some } h, h' \in B(h_t, 2\beta)\}.$$

Intuitively, the disagreement coefficient captures how the measure of this set grows with β. The following is a slight variation of the original definition of Hanneke (2007).

Definition 2. *The* disagreement coefficient $\theta = \theta(\mathcal{D}, \mathcal{H}, \varepsilon) > 0$ *is*

$$\inf_{r \geq \varepsilon + \nu} \left\{ \frac{\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[\exists h \in B(h^*, r) \text{ s.t. } h(x) \neq h^*(x)]}{r} \right\}$$

where $h^* = \arg\inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$ and $\nu = \mathrm{err}_{\mathcal{D}}(h^*)$.

Clearly, $\theta \leq 1/(\varepsilon + \nu)$; furthermore, it is a constant bounded independently of $1/(\varepsilon + \nu)$ in several cases previously considered in the literature (Hanneke, 2007). For example, if $\mathcal{H}$ is homogeneous linear separators and $\mathcal{D}_{\mathcal{X}}$ is the uniform distribution over the unit sphere in $\mathbb{R}^d$, then $\theta = \Theta(\sqrt{d})$.

Algorithm 1
Input: stream $(x_1, x_2, \ldots, x_m)$ i.i.d. from $\mathcal{D}_\mathcal{X}$
Initially, $\hat{S}_0 := \emptyset$ and $T_0 := \emptyset$.
For $n = 1, 2, \ldots, m$:

1. For each $\hat{y} \in \{\pm 1\}$:
   let $h_{\hat{y}} := \mathrm{LEARN}_\mathcal{H}(\hat{S}_{n-1} \cup \{(x_n, \hat{y})\}, T_{n-1})$.
2. If $\mathrm{err}(h_{-\hat{y}}, \hat{S}_{n-1} \cup T_{n-1}) - \mathrm{err}(h_{\hat{y}}, \hat{S}_{n-1} \cup T_{n-1}) > \Delta_{n-1}$ (or if no such $h_{-\hat{y}}$ is found) for some $\hat{y} \in \{\pm 1\}$, then $\hat{S}_n := \hat{S}_{n-1} \cup \{(x_n, \hat{y})\}$ and $T_n := T_{n-1}$.
3. Else request y_n; $\hat{S}_n := \hat{S}_{n-1}$ and $T_n := T_{n-1} \cup \{(x_n, y_n)\}$.

Return $h_f = \mathrm{LEARN}_\mathcal{H}(\hat{S}_m, T_m)$.

Figure 1: The agnostic selective sampling algorithm. See (1) for a possible setting for Δ_n.

Agnostic selective sampling

Here we state and analyze our general algorithm for agnostic active learning. The main techniques employed by the algorithm are reductions to a supervised learning task and generalization bounds applied to differences of empirical errors.

A general algorithm for agnostic active learning

Figure 1 states our algorithm in full generality. The input is a stream of m unlabeled examples drawn i.i.d from $\mathcal{D}_\mathcal{X}$; for the time being, m can be thought of as $\tilde{O}((d/\varepsilon)(1 + \nu/\varepsilon))$ where ε is the accuracy parameter.[1]

The algorithm operates by reduction to a special kind of supervised learning that includes hard constraints.

> For $A, B \subset \mathcal{X} \times \{\pm 1\}$, let $\mathrm{LEARN}_\mathcal{H}(A, B)$ denote a supervised learner that returns a hypothesis $h \in \mathcal{H}$ consistent with A, and with minimum error on B. If there is no hypothesis consistent with A, it reports this.

For some simple hypothesis classes like intervals on the line, or rectangles in $\mathbb{R}^2$, it is easy to construct such a learner. For more complex classes like linear separators, the main bottleneck is the hardness of minimizing the $0-1$ loss on B (that is, the hardness of agnostic supervised learning). If a convex upper bound on this loss function is used instead, as in the case of soft-margin support vector machines, it is straightforward to incorporate hard constraints; but at present the rigorous guarantees accompanying our algorithm apply only if $0-1$ loss is used.

Algorithm 1 maintains two sets of labeled examples, $\hat{S}$ and T, each of which is initially empty. Upon receiving x_n, it learns two[2] hypotheses, $h_{\hat{y}} = \mathrm{LEARN}_\mathcal{H}(\hat{S} \cup \{(x_n, \hat{y})\}, T)$ for $\hat{y} \in \{\pm 1\}$, and then compares their empirical errors on $\hat{S} \cup T$. If the difference is large enough, it is possible to infer how h^* labels x_n (as we show in Lemma 3). In this case, the algorithm adds x_n, with this inferred label, to $\hat{S}$. Otherwise, the algorithm requests the label y_n and adds (x_n, y_n) to T. Thus, $\hat{S}$ contains examples with inferred labels consistent with h^*, and T contains examples with their requested labels. Because h^* might err on some examples in T, we just insist that $\mathrm{LEARN}_\mathcal{H}$ find a hypothesis with minimal error on T. Meanwhile, by construction, h^* is consistent with $\hat{S}$ (as we shall see), so we require $\mathrm{LEARN}_\mathcal{H}$ to only consider hypotheses consistent with $\hat{S}$.

Bounds for error differences

We still need to specify Δ_n, the threshold value for error differences that determines whether the algorithm requests a label or not. Intuitively, Δ_n should reflect how closely empirical errors on a sample approximate true errors on the distribution $\mathcal{D}$. Note that our algorithm is modular with respect to the choice of Δ_n, so, for example, it can be customized for a particular input distribution and hypothesis class. Below we provide a simple and adaptive setting that works for any distribution and hypothesis class with finite VC dimension.

The setting of Δ_n can only depend on observable quantities, so we first clarify the distinction between empirical errors on $\hat{S}_n \cup T_n$ and those with respect to the true (hidden) labels.

Definition 3. *Let $\hat{S}_n$ and T_n be as defined in Algorithm 1. Let S_n (shedding the hat accent) be the set of labeled examples identical to those in $\hat{S}_n$, except with the true hidden labels swapped in. Thus, for example, $S_n \cup T_n$ is an i.i.d. sample from $\mathcal{D}$ of size n. Finally, let*

$$\mathrm{err}_n(h) = \mathrm{err}(h, S_n \cup T_n) \quad \textit{and} \quad \widehat{\mathrm{err}}_n(h) = \mathrm{err}(h, \hat{S}_n \cup T_n).$$

It is straightforward to apply Lemma 1 to empirical errors on $S_n \cup T_n$, i.e. to $\mathrm{err}_n(h)$, but we cannot use such bounds algorithmically: we do not request the true labels for points in $\hat{S}_n$ and thus cannot reliably compute $\mathrm{err}_n(h)$. What we can compute are error *differences* $\mathrm{err}_n(h) - \mathrm{err}_n(h')$ for pairs of hypotheses (h, h') that agree on (and thus make the same mistakes on) $\hat{S}_n$, since for such pairs, we have

$$\mathrm{err}_n(h) - \mathrm{err}_n(h') = \widehat{\mathrm{err}}_n(h) - \widehat{\mathrm{err}}_n(h').^3$$

These empirical error differences are means of $\{-1, 0, +1\}$-valued random variables. We need to rewrite them in terms of $\{0, 1\}$-valued random variables for some of the concentration bounds we will be using.

[1] The $\tilde{O}$ notation suppresses $\log 1/\delta$ and terms polylogarithmic in those that appear.

[2] If $\mathrm{LEARN}_\mathcal{H}$ cannot find a hypothesis consistent with $\hat{S} \cup \{(x_n, y)\}$ for some y, then assuming h^* is consistent with $\hat{S}$, it must be that $h^*(x) = -y$. In this case, we simply add $(x_n, -y)$ to $\hat{S}$, regardless of the error difference.

[3] This observation is enough to immediately justify the use of *additive* generalization bounds for Δ_n. However, we need to use *normalized (multiplicative)* bounds to achieve a better label complexity.

Definition 4. *For a pair* $(h, h') \in \mathcal{H} \times \mathcal{H}$*, define* $g^+_{h,h'}(x, y) = \mathbb{1}[h(x) \neq y \wedge h'(x) = y]$ *and* $g^-_{h,h'}(x, y) = \mathbb{1}[h(x) = y \wedge h'(x) \neq y]$.

With this notation, we have $\mathrm{err}(h, Z) - \mathrm{err}(h', Z) = \mathbb{E}_Z[g^+_{h,h'}] - \mathbb{E}_Z[g^-_{h,h'}]$ for any $Z \subset \mathcal{X} \times \{\pm 1\}$. Now, applying Lemma 1 to $\mathcal{G} = \{g^+_{h,h'} : (h, h') \in \mathcal{H} \times \mathcal{H}\} = \{g^-_{h,h'} : (h, h') \in \mathcal{H} \times \mathcal{H}\}$, and noting that $\mathcal{S}(\mathcal{G}, n) \leq \mathcal{S}(\mathcal{H}, n)^2$, gives the following lemma.

Lemma 2. *Let* $\alpha_n = \sqrt{(4/n)\ln(8\mathcal{S}(\mathcal{H}, 2n)^2/\delta)}$*. With probability at least* $1 - \delta$ *over an i.i.d. sample* Z *of size* n *from* $\mathcal{D}$*, we have for all* $(h, h') \in \mathcal{H} \times \mathcal{H}$*,*

$$\mathrm{err}(h, Z) - \mathrm{err}(h', Z) \leq$$
$$\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathcal{D}}(h') + \alpha_n^2 + \alpha_n\left(\sqrt{\mathbb{E}_Z[g^+_{h,h'}]} + \sqrt{\mathbb{E}_Z[g^-_{h,h'}]}\right).$$

With $Z = S_n \cup T_n$, the error difference on the left-hand side is $\mathrm{err}_n(h) - \mathrm{err}_n(h')$, which can be empirically determined because it is equal to $\widehat{\mathrm{err}}_n(h) - \widehat{\mathrm{err}}_n(h')$. But the terms in the square root on the right-hand side still pose a problem, which we fix next.

Corollary 1. *Let* $\beta_n = \sqrt{(4/n)\ln(8(n^2+n)\mathcal{S}(\mathcal{H}, 2n)^2/\delta)}$*. Then, with probability at least* $1 - \delta$*, for all* $n \geq 1$ *and all* $(h, h') \in \mathcal{H} \times \mathcal{H}$ *consistent with* $\hat{S}_n$*, we have*

$$\widehat{\mathrm{err}}_n(h) - \widehat{\mathrm{err}}_n(h') \leq$$
$$\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathcal{D}}(h') + \beta_n^2 + \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h)} + \sqrt{\widehat{\mathrm{err}}_n(h')}\right).$$

Proof. For each $n \geq 1$, we apply Lemma 2 using $Z = S_n \cup T_n$ and $\delta = \delta/(n^2 + n)$. Then, we apply a union bound over all $n \geq 1$. Thus, with probability at least $1 - \delta$, the bounds in Lemma 2 hold simultaneously for all $n \geq 1$ and all $(h, h') \in \mathcal{H}^2$ with $S_n \cup T_n$ in place of Z. The corollary follows because $\mathrm{err}_n(h) - \mathrm{err}_n(h') = \widehat{\mathrm{err}}_n(h) - \widehat{\mathrm{err}}_n(h')$; and because $\mathbb{E}_{S_n \cup T_n}[g^+_{h,h'}] \leq \widehat{\mathrm{err}}_n(h)$ and $\mathbb{E}_{S_n \cup T_n}[g^-_{h,h'}] \leq \widehat{\mathrm{err}}_n(h')$. To see the first of these expectation bounds, witness that because h and h' agree on S_n,

$$\begin{aligned}
&\mathbb{E}_{S_n \cup T_n}[g^+_{h,h'}] \\
&\quad = \frac{1}{n} \sum_{(x,y) \in T_n} \mathbb{1}[h(x) \neq y \wedge h'(x) = y] \\
&\quad \leq \frac{1}{n} \sum_{(x,y) \in T_n} \mathbb{1}[h(x) \neq y] \;=\; \widehat{\mathrm{err}}_n(h).
\end{aligned}$$

The second bound is similar. □

Corollary 1 implies that we can effectively apply the normalized uniform convergence bounds from Lemma 1 to empirical error differences on $\hat{S}_n \cup T_n$*, even though* $\hat{S}_n \cup T_n$ *is not an i.i.d. sample from* $\mathcal{D}$. In light of this, we use the following setting of Δ_n:

$$\Delta_n := \beta_n^2 + \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h_{+1})} + \sqrt{\widehat{\mathrm{err}}_n(h_{-1})}\right) \qquad (1)$$

where $\beta_n = \sqrt{(4/n)\ln(8(n^2+n)\mathcal{S}(\mathcal{H}, 2n)^2/\delta)} = \tilde{O}(\sqrt{d \log n / n})$ as per Corollary 1.

Correctness and fall-back analysis

We now justify our setting of Δ_n with a correctness proof and fall-back guarantee.

The following lemma elucidates how the inferred labels in $\hat{S}$ serve as a mechanism for implicitly maintaining a candidate set of hypotheses that always includes h^*. The fall-back guarantee then follows almost immediately.

Lemma 3. *With probability at least* $1 - \delta$*, the hypothesis* $h^* = \arg\inf_{h \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$ *is consistent with* $\hat{S}_n$ *for all* $n \geq 0$ *in Algorithm 1.*

Proof. Apply the bounds in Corollary 1 (they hold with probability at least $1 - \delta$) and proceed by induction on n. The base case is trivial since $\hat{S}_0 = \emptyset$. Now assume h^* is consistent with $\hat{S}_n$. Suppose upon receiving x_{n+1}, we discover $\widehat{\mathrm{err}}_n(h_{+1}) - \widehat{\mathrm{err}}_n(h_{-1}) > \Delta_n$. We will show that $h^*(x_{n+1}) = -1$ (assume both h_{+1} and h_{-1} exist, since it is clear $h^*(x_{n+1}) = -1$ if h_{+1} does not exist). Suppose for the sake of contradiction that $h^*(x_{n+1}) = +1$. We know that $\widehat{\mathrm{err}}_n(h^*) \geq \widehat{\mathrm{err}}_n(h_{+1})$ (by the inductive hypothesis h^* is consistent with $\hat{S}_n$, and yet the learner chose h_{+1} in preference to it) and $\widehat{\mathrm{err}}_n(h_{+1}) - \widehat{\mathrm{err}}_n(h_{-1}) > \beta_n^2 + \beta_n(\sqrt{\widehat{\mathrm{err}}_n(h_{+1})} + \sqrt{\widehat{\mathrm{err}}_n(h_{-1})})$. In particular, $\widehat{\mathrm{err}}_n(h_{+1}) > \beta_n^2$. Therefore,

$$\begin{aligned}
&\widehat{\mathrm{err}}_n(h^*) - \widehat{\mathrm{err}}_n(h_{-1}) \\
&= (\widehat{\mathrm{err}}_n(h^*) - \widehat{\mathrm{err}}_n(h_{+1})) + (\widehat{\mathrm{err}}_n(h_{+1}) - \widehat{\mathrm{err}}_n(h_{-1})) \\
&> \sqrt{\widehat{\mathrm{err}}_n(h_{+1})}\left(\sqrt{\widehat{\mathrm{err}}_n(h^*)} - \sqrt{\widehat{\mathrm{err}}_n(h_{+1})}\right) \\
&\quad + \beta_n^2 + \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h_{+1})} + \sqrt{\widehat{\mathrm{err}}_n(h_{-1})}\right) \\
&> \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h^*)} - \sqrt{\widehat{\mathrm{err}}_n(h_{+1})}\right) \\
&\quad + \beta_n^2 + \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h_{+1})} + \sqrt{\widehat{\mathrm{err}}_n(h_{-1})}\right) \\
&= \beta_n^2 + \beta_n\left(\sqrt{\widehat{\mathrm{err}}_n(h^*)} + \sqrt{\widehat{\mathrm{err}}_n(h_{-1})}\right).
\end{aligned}$$

Now Corollary 1 implies that $\mathrm{err}_{\mathcal{D}}(h^*) > \mathrm{err}_{\mathcal{D}}(h_{-1})$, a contradiction. □

Theorem 1. *Let* $\nu = \inf_{h \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$ *and* $d = \mathrm{vcdim}(\mathcal{H})$*. There exists a constant* $c > 0$ *such that the following holds. If Algorithm 1 is given a stream of* m *unlabeled examples, then with probability at least* $1 - \delta$*, the algorithm returns a hypothesis with error at most* $\nu + c \cdot ((1/m)(d \log m + \log(1/\delta)) + \sqrt{(\nu/m)(d \log m + \log(1/\delta))})$.

Proof. Lemma 3 implies that h^* is consistent with $\hat{S}_m$ with probability at least $1 - \delta$. Using the same bounds from Corollary 1 (already applied in Lemma 3) on h^* and h_f together with the fact $\widehat{\mathrm{err}}_m(h_f) \leq \widehat{\mathrm{err}}_m(h^*)$, we have $\mathrm{err}_{\mathcal{D}}(h_f) \leq \nu + \beta_m^2 + \beta_m\sqrt{\nu} + \beta_m\sqrt{\mathrm{err}_{\mathcal{D}}(h_f)}$, which in turn implies $\mathrm{err}_{\mathcal{D}}(h_f) \leq \nu + 3\beta_m^2 + 2\beta_m\sqrt{\nu}$. □

So, Algorithm 1 returns a hypothesis with error at most $\nu + \varepsilon$ when $m = \tilde{O}((d/\varepsilon)(1 + \nu/\varepsilon))$; this is (asymptotically) the usual sample complexity of supervised learning. Since the

algorithm requests at most m labels, its label complexity is always at most $\tilde{O}((d/\varepsilon)(1+\nu/\varepsilon))$.

Label complexity analysis

We can also bound the label complexity of our algorithm in terms of the disagreement coefficient θ. This yields tighter bounds when θ is bounded independently of $1/(\varepsilon+\nu)$. The key to deriving our label complexity bounds based on θ is noting that the probability of requesting the $(n+1)$st label is intimately related to θ and Δ_n.

Lemma 4. *There exist constants $c_1, c_2 > 0$ such that, with probability at least $1-2\delta$, for all $n \geq 1$, the following holds. Let $h^*(x_{n+1}) = \hat{y}$ where $h^* = \arg\inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$. Then, the probability that Algorithm 1 requests the label y_{n+1} is*

$$\begin{aligned}&\Pr_{x_{n+1}\sim\mathcal{D}_{\mathcal{X}}}[\textit{Request } y_{n+1}]\\ &\quad\leq \Pr_{x_{n+1}\sim\mathcal{D}_{\mathcal{X}}}[\mathrm{err}_{\mathcal{D}}(h_{-\hat{y}}) \leq c_1\nu + c_2\beta_n^2]\end{aligned}$$

where β_n is as defined in Corollary 1 and $\nu = \inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$.

Proof. See full version of paper. □

Lemma 5. *In the same setting as Lemma 4, there exists a constant $c > 0$ such that $\Pr_{x_{n+1}\sim\mathcal{D}_{\mathcal{X}}}[\textit{Request } y_{n+1}] \leq c \cdot \theta \cdot (\nu + \beta_n^2)$, where $\theta = \theta(\mathcal{D}, \mathcal{H}, 3\beta_m^2 + 2\beta_m\sqrt{\nu})$ is the disagreement coefficient, $\nu = \inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$, and β_n is as defined in Corollary 1.*

Proof. Suppose $h^*(x_{n+1}) = -1$. By the triangle inequality, we have that $\mathrm{err}_{\mathcal{D}}(h_{+1}) \geq \rho(h_{+1}, h^*) - \nu$, where ρ is the disagreement metric on $\mathcal{H}$ (Definition 1). By Lemma 4, this implies that the probability of requesting y_{n+1} is at most the probability that $\rho(h_{+1}, h^*) \leq (c_1+1)\nu + c_2\beta_n^2$ for some constants $c_1, c_2 > 0$. We can choose the constants so that $(c_1+1)\nu + c_2\beta_n^2 \geq \nu + 3\beta_m^2 + 2\beta_m\sqrt{\nu}$. Then, the definition of the disagreement coefficient gives the conclusion that $\Pr_{x_{n+1}\sim\mathcal{D}_{\mathcal{X}}}[\rho(h_{+1}, h^*) \leq (c_1+1)\nu + c_2\beta_n^2] \leq \theta \cdot ((c_1+1)\nu + c_2\beta_n^2)$. □

Now we give our main label complexity bound for agnostic active learning.

Theorem 2. *Let m be the number of unlabeled data given to Algorithm 1, $d = \mathrm{vcdim}(\mathcal{H})$, $\nu = \inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$, β_m as defined in Corollary 1, and $\theta = \theta(\mathcal{D}, \mathcal{H}, 3\beta_m^2 + 2\beta_m\sqrt{\nu})$. There exists a constant $c_1 > 0$ such that for any $c_2 \geq 1$, with probability at least $1-2\delta$:*

1. *If $\nu \leq (c_2-1)\beta_m^2$, Algorithm 1 returns a hypothesis with error as bounded in Theorem 1 and the expected number of labels requested is at most*

$$1 + c_1c_2\theta \cdot \left(d\log^2 m + \log\frac{1}{\delta}\log m\right).$$

2. *Else, the same holds except the expected number of labels requested is at most*

$$1 + c_1\theta \cdot \left(\nu m + d\log^2 m + \log\frac{1}{\delta}\log m\right).$$

Furthermore, if L is the expected number of labels requested as per above, then with probability at least $1-\delta'$, the algorithm requests no more than $L + \sqrt{3L\log(1/\delta')}$ labels.

Proof. Follows from Lemma 5 and a Chernoff bound for the Poisson trials $\mathbb{1}[\text{Request } y_n]$. □

With the substitution $\varepsilon = 3\beta_m^2 + 2\beta_m\sqrt{\nu}$ as per Theorem 1, Theorem 2 entails that for any hypothesis class and data distribution for which the disagreement coefficient $\theta = \theta(\mathcal{D}, \mathcal{H}, \varepsilon)$ is bounded independently of $1/(\varepsilon+\nu)$ (see (Hanneke, 2007) for some examples), Algorithm 1 only needs $\tilde{O}(\theta d\log^2(1/\varepsilon))$ labels to achieve error $\varepsilon \approx \nu$ and $\tilde{O}(\theta d(\log^2(1/\varepsilon) + (\nu/\varepsilon)^2))$ labels to achieve error $\varepsilon \ll \nu$. The latter matches the dependence on ν/ε in the $\Omega((\nu/\varepsilon)^2)$ lower bound (Kääriäinen, 2006).

The linear dependence on θ improves on the quadratic dependence shown for A^2 (Hanneke, 2007)[4]. For an illustrative consequence of this, suppose $\mathcal{D}_{\mathcal{X}}$ is the uniform distribution on the sphere in $\mathbb{R}^d$ and $\mathcal{H}$ is homogeneous linear separators; in this case, $\theta = \Theta(\sqrt{d})$. Then the label complexity of A^2 depends at least quadratically on the dimension, whereas the corresponding dependence for our algorithm is $d^{3/2}$. A specially-designed setting of Δ_n (say, specific to the input distribution and hypothesis class) may be able to further reduce the dependence to d (see Balcan, Broder, and Zhang (2007)).

Experiments

We implemented Algorithm 1 in a few simple cases to experimentally demonstrate the label complexity improvements. In each case, the data distribution $\mathcal{D}_{\mathcal{X}}$ was uniform over $[0,1]$; the stream length was $m = 10000$, and each experiment was repeated 20 times with different random seeds. Our first experiment studied linear thresholds on the line. The target hypothesis was fixed to be $h^*(x) = \mathrm{sign}(x-0.5)$. For this hypothesis class, we used two different noise models, each of which ensured $\inf_{h\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h) = \mathrm{err}_{\mathcal{D}}(h^*) = \nu$ for a pre-specified $\nu \in [0,1]$. The first model was random misclassification: for each point $x \sim \mathcal{D}_{\mathcal{X}}$, we independently labeled it $h^*(x)$ with probability $1-\nu$ and $-h^*(x)$ with probability ν. In the second model (also used in Castro and Nowak (2006)), for each point $x \sim \mathcal{D}_{\mathcal{X}}$, we independently labeled it $+1$ with probability $(x-0.5)/(4\nu)+0.5$ and -1 otherwise, thus concentrating the noise near the boundary. Our second experiment studied intervals on the line. Here, we only used random misclassification, but we varied the target interval length $p_+ = \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[h^*(x) = +1]$.

[4] It may be possible to reduce A^2's quadratic dependence to a linear dependence by using normalized bounds, as we do here.

The results show that the number of labels requested by Algorithm 1 was exponentially smaller than the total number of data seen (m) under the first noise model, and was polynomially smaller under the second noise model (see Figure 2; we verified the polynomial vs. exponential distinction on separate log-log scale plots). In the case of intervals, we observe an initial phase (of duration roughly $\propto 1/p_+$) in which every label is requested, followed by a more efficient phase, confirming the known active-learnability of this class. These improvements show that our algorithm needed significantly fewer labels to achieve the same error as a standard supervised algorithm that uses labels for all points seen.

As a sanity check, we examined the locations of data for which Algorithm 1 requested a label. We looked at two particular runs of the algorithm: the first was with $\mathcal{H} =$ intervals, $p_+ = 0.2$, $m = 10000$, and $\nu = 0.1$; the second was with $\mathcal{H} =$ boxes ($d = 2$), $p_+ = 0.49$, $m = 1000$, and $\nu = 0.01$. In each case, the data distribution was uniform over $[0, 1]^d$, and the noise model was random misclassification. Figure 3 shows that, early on, labels were requested everywhere. But as the algorithm progressed, label requests concentrated near the boundary of the target hypothesis.

Conclusion and future work

We have presented a simple and natural approach to agnostic active learning. Our extension of the selective sampling scheme of Cohn-Atlas-Ladner (1994)

1. simplifies the maintenance of the region of uncertainty with a reduction to supervised learning, and
2. guards against noise with a suitable algorithmic application of generalization bounds.

Our algorithm relies on a threshold parameter Δ_n for comparing empirical errors. We prescribe a very simple and natural choice for Δ_n – a normalized generalization bound from supervised learning – but one could hope for a more clever or aggressive choice, akin to those in Balcan, Broder, and Zhang (2007) for linear separators.

Finding consistent hypotheses when data is separable is often a simple task. In such cases, reduction-based active learning algorithms can be relatively efficient. On the other hand, agnostic supervised learning is computationally intractable for many hypothesis classes (Guruswami and Raghavendra, 2006), and of course, agnostic active learning is at least as hard in the worst case. Our reduction to supervised learning is benign in the sense that the learning problems we need to solve are over samples from the original distribution, so we do not create pathologically hard instances (like those arising from hardness reductions) unless they are inherent in the data. Nevertheless, an important research direction is to develop consistent active learning algorithms that only require solving tractable (e.g. convex) optimization problems. A similar reduction-based scheme may be possible.

Acknowledgements

We are grateful to the Engineering Institute (a research and educational partnership between Los Alamos National Laboratory and U.C. San Diego) for supporting the second author with a graduate fellowship, and to the NSF for support under grants IIS-0347646 and IIS-0713540.

References

Balcan, M.-F.; Beygelzimer, A.; and Langford, J. 2006. Agnostic active learning. In *ICML*.

Balcan, M.-F.; Broder, A.; and Zhang, T. 2007. Margin based active learning. In *COLT*.

Bousquet, O.; Boucheron, S.; and Lugosi, G. 2004. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence* 3176:169–207.

Castro, R., and Nowak, R. 2006. Upper and lower bounds for active learning. In *Allerton Conference on Communication, Control and Computing*.

Castro, R., and Nowak, R. 2007. Minimax bounds for active learning. In *COLT*.

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

Dasgupta, S.; Kalai, A.; and Monteleoni, C. 2005. Analysis of perceptron-based active learning. In *COLT*.

Dasgupta, S. 2005. Coarse sample complexity bounds for active learning. In *NIPS*.

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

Gilad-Bachrach, R.; Navot, A.; and Tishby, N. 2005. Query by committeee made real. In *NIPS*.

Guruswami, V., and Raghavendra, P. 2006. Hardness of learning halfspaces with noise. In *FOCS*.

Hanneke, S. 2007a. A bound on the label complexity of agnostic active learning. In *ICML*.

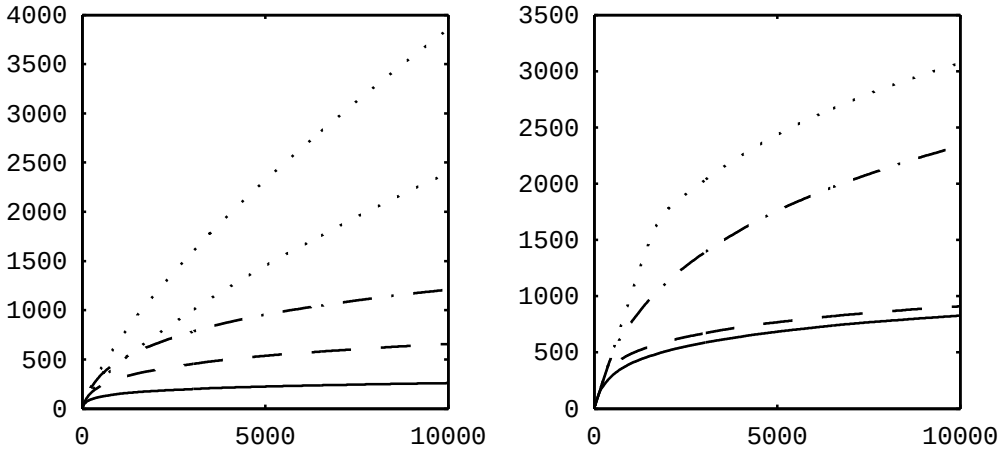
Hanneke, S. 2007b. Teaching dimension and the complexity of active learning. In *COLT*.

Kääriäinen, M. 2006. Active learning in the non-realizable case. In *ALT*.

Monteleoni, C. 2006a. Efficient algorithms for general active learning. In *COLT*. Open problem.

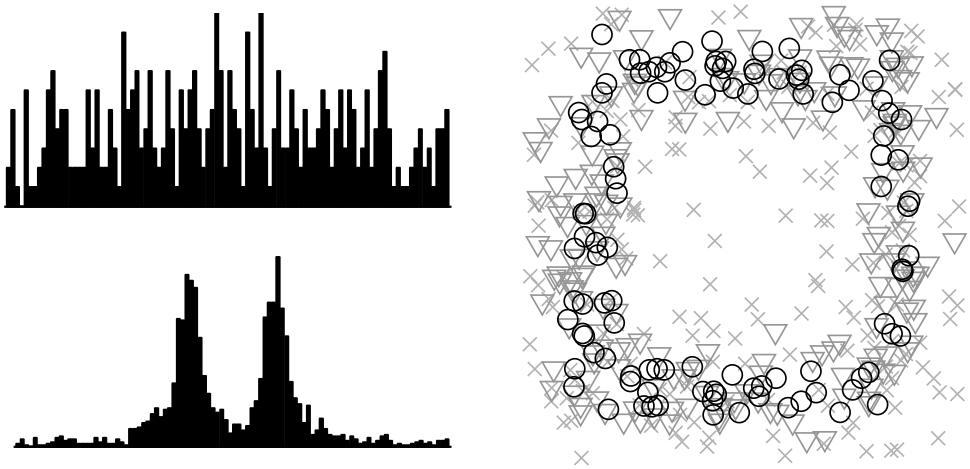
Monteleoni, C. 2006b. *Learning with online constraints: shifting concepts and active learning*. PhD Thesis, MIT Computer Science and Artificial Intelligence Laboratory.

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



(a) (b)

Figure 2: Labeling rate plots. The plots show the number of labels requested (vertical axis) versus the total number of points seen (labeled + unlabeled, horizontal axis) using Algorithm 1. (a) $\mathcal{H}$ = thresholds: under random misclassification noise with $\nu = 0$ (solid), 0.1 (dashed), 0.2 (dot-dashed); under the boundary noise model with $\nu = 0.1$ (lower dotted), 0.2 (upper dotted). (b) $\mathcal{H}$ = intervals: under random misclassification with $(p_+, \nu) = (0.2, 0.0)$ (solid), $(0.1, 0.0)$ (dashed), $(0.2, 0.1)$ (dot-dashed), $(0.1, 0.1)$ (dotted).



(a) (b)

Figure 3: Locations of label requests. (a) $\mathcal{H}$ = intervals, $h^* = [0.4, 0.6]$. The top histogram shows the locations of first 400 label requests (the x-axis is the unit interval); the bottom histogram is for all (2141) label requests. (b) $\mathcal{H}$ = boxes, $h^* = [0.15, 0.85]^2$. The first 200 requests occurred at the $\times$s, the next 200 at the $\bigtriangledown$s, and the final 109 at the $\bigcirc$s.