Importance Weighted Active Learning

Alina Beygelzimer BEYGEL@US.IBM.COM
IBM Research, 19 Skyline Drive, Hawthorne, NY 10532

Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
Department of Computer Science and Engineering, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093

John Langford JL@YAHOO-INC.COM
Yahoo! Research, 111 West 40th Street, New York, NY 10018

Abstract

We present a practical and statistically consistent scheme for actively learning binary classifiers under general loss functions. Our algorithm uses importance weighting to correct sampling bias, and by controlling the variance, we are able to give rigorous label complexity bounds for the learning process.

1. Introduction

An active learner interactively chooses which data points to label, while a passive learner obtains all the labels at once. The great hope of active learning is that interaction can substantially reduce the number of labels required, making learning more practical. This hope is known to be valid in certain special cases, where the number of label queries has been shown to be logarithmic in the usual sample complexity of passive learning; such cases include thresholds on a line, and linear separators with a spherically uniform unlabeled data distribution (Dasgupta et al., 2005).

Many earlier active learning algorithms, such as (Cohn et al., 1994; Dasgupta et al., 2005), are not consistent when data is not perfectly separable under the given hypothesis class: even with an infinite labeling budget, they might not converge to an optimal predictor (see Dasgupta and Hsu (2008) for a discussion).

This problem has recently been addressed in two threads of research. One approach (Balcan et al., 2006; Dasgupta et al., 2008; Hanneke, 2007) constructs learning algorithms that explicitly use sample complexity bounds to assess which hypotheses are still "in the running," and thereby assess the relative value of different unlabeled points (in terms of whether they help distinguish between the remaining hypotheses). These algorithms have the usual PAC-style convergence guarantees, but also have rigorous label complexity bounds that are in many cases significantly better than the bounds for passive learning. The second approach to active learning uses importance weights to correct sampling bias (Bach, 2007; Sugiyama, 2006).

The PAC-guarantee active learning algorithms have yet to see practical use for several reasons. First, they are built explicitly for 0–1 loss and are not easily adapted to most other loss functions. This is problematic because in many applications, other loss functions are more appropriate for describing the problem, or make learning more tractable (as with convex proxy losses on linear representations). Second, these algorithms make internal use of generalization bounds that are often loose in practice, and they can thus end up requiring far more labels than are really necessary. Finally, they typically require an explicit enumeration over the hypothesis class (or an ϵ-cover thereof), which is generally computationally intractable.

Importance weighted approaches have only been analyzed in limited settings. For example, Bach (2007) considers linear models and provides an analysis of consistency in cases where either (i) the model class fits the data perfectly, or (ii) the sampling strategy is non-adaptive (that is, a query doesn't depend on the sequence of previous queries). Also, the analysis in these works is asymptotic rather than yielding finite label bounds. Label complexity is of paramount importance in active learning, because otherwise simpler passive learning approaches can be used. Furthermore,

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

a poor choice of importance weights can result in high label complexity.

We address the problems above with a new active learning scheme that provably yields PAC-style label complexity guarantees. When presented with an unlabeled point x_t, this scheme queries its label with a carefully chosen probability p_t, taking into account the identity of the point and the history of labels seen so far. The points that end up getting labeled are then weighted according to the reciprocals of these probabilities (that is, $1/p_t$), in order to remove sampling bias. We show (theorem 1) that this simple method guarantees statistical consistency: for any distribution and any hypothesis class, active learning eventually converges to the optimal hypothesis in the class.

As in any importance sampling scenario, the biggest challenge is controlling the variance of the process. This depends crucially on how the sampling probability p_t is chosen. Roughly, our strategy is to make it proportional to the spread of values $h(x_t)$, as h ranges over the remaining candidate hypotheses. For this setting of p_t, which we call loss-weighting, we have two results: a fallback guarantee that the label complexity is never much worse than that of supervised learning (theorem 2), and a rigorous label complexity bound (theorem 11). Previously, label complexity bounds for active learning were only known for 0–1 loss, and were based on the *disagreement coefficient* of the learning problem (Hanneke, 2007). We generalize this notion to general loss functions, and analyze label complexity in terms of it. We consider settings in which these bounds turn out to be roughly the *square root* of the sample complexity of supervised learning.

In addition to these upper bounds, we show a general lower bound on the label complexity of active learning (theorem 12) that significantly improves the best previous such result (Kääriäinen, 2006).

We conduct practical experiments with two IWAL algorithms. The first is a specialization of IWAL with loss-weighting to the case of linear classifiers with convex loss functions; here, the algorithm becomes tractable via convex programming (section 7). The second uses a simple bootstrapping scheme that reduces active learning to passive learning without requiring much additional computation (section 7.2). In every case, these experiments yield substantial reductions in label complexity compared to passive learning, without compromising predictive performance. They suggest that IWAL is a practical scheme that can reduce the label complexity of active learning without sacrificing the statistical guarantees (like consistency) we take for granted in passive learning.

Boosting and bagging-based algorithms of Abe and Mamitsuka (1998) are similar in spirit to our bootstrapping IWAL scheme in section 7.2, but they are not consistent in the presence of noise.

2. Preliminaries

Let X be the input space and Y the output space. We consider active learning in the streaming setting where at each step t, a learner observes an unlabeled point $x_t \in X$ and has to decide whether to ask for the label $y_t \in Y$. The learner works with a hypothesis space $H = \{h : X \to Z\}$, where Z is a prediction space.

The algorithm is evaluated with respect to a given loss function $l : Z \times Y \to [0, \infty)$. The most common loss function is 0–1 loss, in which $Y = Z = \{-1, 1\}$ and $l(z, y) = \mathbf{1}(y \neq z) = \mathbf{1}(yz < 0)$. The following examples address the binary case $Y = \{-1, 1\}$ with $Z \subset \mathbb{R}$: $l(z, y) = (1 - yz)_+$ (hinge loss), $l(z, y) = \ln(1 + e^{-yz})$ (logistic loss), $l(z, y) = (y - z)^2 = (1 - yz)^2$ (squared loss), and $l(z, y) = |y - z| = |1 - yz|$ (absolute loss). Notice that all the loss functions mentioned here are of the form $l(z, y) = \phi(yz)$ for some function ϕ on the reals. We specifically highlight this subclass of loss functions when proving label complexity bounds. Since these functions are bounded (if Z is), we further assume they are normalized to output a value in $[0, 1]$.

3. The Importance Weighting Skeleton

Algorithm 1 describes the basic outline of importance-weighted active learning (IWAL). Upon seeing x_t, the learner calls a subroutine *rejection-threshold* (instantiated in later sections), which looks at x_t and past history to return the probability p_t of requesting y_t. The algorithm maintains a set of labeled examples seen so far, each with an importance weight: if y_t ends up being queried, its weight is set to $1/p_t$.

Algorithm 1 IWAL (subroutine rejection-threshold)

Set $S_0 = \emptyset$.

For t from $1, 2, ...$ until the data stream runs out:

1. Receive x_t .
2. Set p_t = rejection-threshold$(x_t, \{x_i, y_i, p_i, Q_i : 1 \leq i < t\})$.
3. Flip a coin $Q_t \in \{0, 1\}$ with $\mathbf{E}[Q_t] = p_t$. If $Q_t = 1$, request y_t and set $S_t = S_{t-1} \cup \{(x_t, y_t, 1/p_t)\}$, else $S_t = S_{t-1}$.
4. Let $h_t = \arg\min_{h \in H} \sum_{(x,y,c) \in S_t} c \cdot l(h(x), y)$.

Examples are assumed to be drawn *i.i.d.* from the un-

derlying probability distribution D. The expected loss of $h \in H$ on D is given by $L(h) = \mathbf{E}_{(x,y)\sim D}\, l(h(x), y)$. The importance weighted estimate of the loss at time T is

$$L_T(h) = \frac{1}{T}\sum_{t=1}^{T} \frac{Q_t}{p_t}\, l(h(x_t), y_t),$$

where Q_t is as defined in the algorithm. It is not hard to see that $\mathbf{E}[L_T(h)] = L(h)$, with the expectation taken over all the random variables involved. Theorem 2 gives large deviation bounds for $L_T(h)$, provided that the probabilities p_t are chosen carefully. Even though step 4 is stated as empirical risk minimization, any passive importance-weighted learning algorithm can be used to learn h_t based on S_t.

3.1. A safety guarantee for IWAL

A desirable property for a learning algorithm is *consistency*: Given an infinite budget of unlabeled and labeled examples, does the algorithm converge to the best predictor? Some early active learning algorithms (Cohn et al., 1994; Dasgupta et al., 2005) do not satisfy this baseline guarantee if the points cannot be classified perfectly by the given hypothesis class. We prove that IWAL algorithms are consistent, as long as p_t is bounded away from 0.

Comparing this result to the usual sample complexity bounds in supervised learning (for example, corollary 4.2 of (Langford, 2005)), we see that the label complexity is at most $2/p_{\min}^2$ times that of a supervised algorithm. For simplicity, the bound is given in terms of $\ln|H|$ rather than the VC dimension of H. The argument, which is a martingale modification of standard results, can be extended to VC spaces.

Theorem 1. *For all distributions D, for all finite hypothesis classes H, for any $\delta > 0$, if there is a constant $p_{\min} > 0$ such that $p_t \geq p_{\min}$ for all $1 \leq t \leq T$, then*

$$\mathbf{P}\left[\max_{h\in H} |L_T(h) - L(h)| > \frac{\sqrt{2}}{p_{\min}}\sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{T}}\right] < \delta.$$

Proof. Fix D. For a hypothesis $h \in H$, consider a sequence of random variables $U_1, \ldots, U_T$ with $U_t = \frac{Q_t}{p_t} l(h(x_t), y_t) - L(h)$. Since $p_t \geq p_{\min}$, $|U_t| \leq 1/p_{\min}$. The sequence $Z_t = \sum_{i=1}^{t} U_i$ is a martingale, letting $Z_0 = 0$. Indeed, for any $1 \leq t \leq T$,

$$\mathbf{E}[Z_t \mid Z_{t-1}, ..., Z_0] = \mathbf{E}\left[U_t + Z_{t-1} \mid Z_{t-1}, ..., Z_0\right]$$
$$= \mathbf{E}\left[\, l(h(x_t), y_t) - L(h) + Z_{t-1} \mid Z_{t-1}, ..., Z_0 \,\right] = Z_{t-1}.$$

Observe that $|Z_{t+1} - Z_t| = |U_{t+1}| \leq 1/p_{\min}$ for all $0 \leq t < T$. Using $Z_T = T(L_T(h) - L(h))$ and applying Azuma's inequality (1967), we see that for any $\lambda > 0$,

$$\mathbf{P}\left[|L_T(h) - L(h)| > \frac{\lambda}{p_{\min}\sqrt{T}}\right] < 2e^{-\lambda^2/2}.$$

Setting $\lambda = \sqrt{2(\ln|H| + \ln(2/\delta))}$ and taking a union bound over $h \in H$ then yields the desired result. □

4. Setting the rejection threshold

Algorithm 2 gives a particular instantiation of the rejection threshold subroutine in IWAL. The subroutine maintains an effective hypothesis class H_t, which is initially all of H and then gradually shrinks by setting H_{t+1} to the subset of H_t whose empirical loss isn't too much worse than L_t^*, the smallest empirical loss in H_t:

$$H_{t+1} = \{h \in H_t : L_t(h) \leq L_t^* + \Delta_t\}.$$

The allowed slack $\Delta_t = \sqrt{(8/t)\ln(2t(t+1)|H|^2/\delta)}$ comes from a standard sample complexity bound.

We will show that, with high probability, any optimal hypothesis h^* is always in H_t, and thus all other hypotheses can be discarded from consideration. For each x_t, the loss-weighting scheme looks at the range of predictions on x_t made by hypotheses in H_t and sets the sampling probability p_t to (roughly) the size of this range: precisely, $p_t = \max_{f,g\in H_t} \max_y l(f(x_t), y) - l(g(x_t), y)$. Since the loss values are normalized to lie in $[0, 1]$, we can be sure that p_t is also in this interval.

Algorithm 2 loss-weighting $(x, \{x_i, y_i, p_i, Q_i : i < t\})$

1. Initialize $H_0 = H$.
2. Update $L_{t-1}^* = \min_{h\in H_{t-1}} \frac{1}{t-1}\sum_{i=1}^{t-1} \frac{Q_i}{p_i} l(h(x_i), y_i)$
   Set H_t to
   $$\left\{h \in H_{t-1} : \frac{1}{t-1}\sum_{i=1}^{t-1} \frac{Q_i}{p_i} l(h(x_i), y_i) \leq L_{t-1}^* + \Delta_{t-1}\right\}$$
3. Return $p_t = \max_{f,g\in H_t, y\in Y} l(f(x), y) - l(g(x), y)$.

4.1. A generalization bound

We start with a large deviation bound for each h_t output by IWAL(loss-weighting). It is not a corollary of theorem 1 because it does not require the sampling probabilities be bounded below away from zero.

Theorem 2. *Pick any learning problem D and hypothesis class H, and let $h^* \in H$ be a minimizer of the loss function with respect to D. For any $\delta > 0$, with probability at least $1 - \delta$, for any T:* (i) *$h^* \in H_T$, and*

(ii) $L(f) - L(g) \leq 2\Delta_{T-1}$ *for any* $f, g \in H_T$. *In particular, if* h_T *is the hypothesis output by* IWAL(loss-weighting), $L(h_T) - L(h^*) \leq 2\Delta_{T-1}$.

We need the following lemma for the proof.

Lemma 3. *For all learning problems* D, *for all hypothesis classes* H, *for all* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* T *and all* $f, g \in H_T$, $|L_T(f) - L_T(g) - L(f) + L(g)| \leq \Delta_T$.

Proof. Pick any T and $f, g \in H_T$, and define $Z_t = \frac{Q_t}{p_t}\big(l(f(x_t), y_t) - l(g(x_t), y_t)\big) - (L(f) - L(g))$. Then

$$\mathbf{E}\left[\, Z_t \mid Z_1, \cdots, Z_{t-1}\right] = \mathbf{E}_{x_t, y_t}[\, l(f(x_t), y_t) - l(g(x_t), y_t) \mid Z_1, \cdots, Z_{t-1}] - (L(f) - L(g)) = 0.$$

Thus $Z_1, Z_2, \ldots$ is a martingale difference sequence. We can use Azuma's inequality to show that its sum is tightly concentrated, if the individual Z_t are bounded.

To check boundedness, observe that since f, g are in H_T, they must also be in $H_1, H_2, \ldots, H_{T-1}$. Thus for all $t \leq T$, $p_t \geq |l(f(x_t), y_t) - l(g(x_t), y_t)|$, whereupon $|Z_t| \leq \frac{1}{p_t}|l(f(x_t), y_t) - l(g(x_t), y_t)| + |L(f) - L(g)| \leq 2$.

We allow failure probability $\delta / 2T(T+1)$ at time T. Applying Azuma's inequality, we have

$$\mathbf{P}[|L_T(f) - L_T(g) - L(f) + L(g)| \geq \Delta_T] = \mathbf{P}\left[\left|\sum_{t=1}^{T} Z_t\right| \geq T\Delta_T\right] \leq 2e^{-T\Delta_T^2/8} = \frac{\delta}{T(T+1)|H|^2}.$$

Since H_T is a random subset of H, it suffices to take a union bound over all $f, g \in H$. A union bound over T finishes the proof. □

Proof of theorem 2. Start by assuming that the $1 - \delta$ probability event of lemma 3 holds. We first show by induction that $h^* = \arg\min_{h \in H} L(h)$ is in H_T for all T. It holds at $T = 1$, since $H_1 = H_0 = H$. Now suppose it holds at T, and show that it is true at $T+1$. Let h_T minimize L_T over H_T. By lemma 3, $L_T(h^*) - L_T(h_T) \leq L(h^*) - L(h_T) + \Delta_T \leq \Delta_T$. Thus $L_T(h^*) \leq L_T^* + \Delta_T$ and hence $h^* \in H_{T+1}$.

Next, we show that for any $f, g \in H_T$ we have $L(f) - L(g) \leq 2\Delta_{T-1}$. By lemma 3, since $H_T \subseteq H_{T-1}$, $L(f) - L(g) \leq L_{T-1}(f) - L_{T-1}(g) + \Delta_{T-1} \leq L_{T-1}^* + \Delta_{T-1} - L_{T-1}^* + \Delta_{T-1} = 2\Delta_{T-1}$. Since $h_T, h^* \in H_T$, we have $L(h_T) \leq L(h^*) + 2\Delta_{T-1}$. □

5. Label Complexity

We showed that the loss of the classifier output by IWAL(loss-weighting) is similar to the loss of the classifier chosen passively after seeing all T labels. How many of those T labels does the active learner request? Dasgupta et al. (2008) studied this question for an active learning scheme under 0–1 loss. For learning problems with bounded *disagreement coefficient* (Hanneke, 2007), the number of queries was found to be $O(\eta T + d \log^2 T)$, where d is the VC dimension of the function class, and η is the best error rate achievable on the underlying distribution by that function class. We will soon see (section 6) that the term ηT is inevitable for any active learning scheme; the remaining term has just a polylogarithmic dependence on T.

We generalize the disagreement coefficient to arbitrary loss functions and show that, under conditions similar to the earlier result, the number of queries is $O\left(\eta T + \sqrt{dT \log^2 T}\right)$, where η is now the best achievable loss. The inevitable ηT is still there, and the second term is still sublinear.

5.1. Label Complexity: Main Issues

Suppose the loss function is minimized by $h^* \in H$, with $L^* = L(h^*)$. Theorem 2 shows that at time t, the remaining hypotheses H_t include h^* and all have losses in the range $[L^*, L^* + 2\Delta_{t-1}]$. We now prove that under suitable conditions, the sampling probability p_t has expected value $\approx L^* + \Delta_{t-1}$. Thus the expected total number of labels queried upto time T is roughly $L^*T + \sum_{t=1}^{T} \Delta_{t-1} \approx L^*T + \sqrt{T \ln |H|}$.

To motivate the proof, consider a loss function $l(z, y) = \phi(yz)$; all our examples are of this form. Say ϕ is differentiable with $0 < C_0 \leq |\phi'| \leq C_1$. Then the sampling probability for x_t is

$$\begin{aligned} p_t &= \max_{f,g \in H_t} \max_{y} l(f(x_t), y) - l(g(x_t), y) \\ &= \max_{f,g \in H_t} \max_{y} \phi(yf(x_t)) - \phi(yg(x_t)) \\ &\leq C_1 \max_{f,g \in H_t} \max_{y} |yf(x_t) - yg(x_t)| \\ &= C_1 \max_{f,g \in H_t} |f(x_t) - g(x_t)| \\ &\leq 2C_1 \max_{h \in H_t} |h(x_t) - h^*(x_t)|. \end{aligned}$$

So p_t is determined by the range of predictions on x_t by hypotheses in H_t. Can we bound the size of this range, given that any $h \in H_t$ has loss at most $L^* + 2\Delta_{t-1}$?

$$\begin{aligned} 2\Delta_{t-1} &\geq L(h) - L^* \\ &\geq \mathbf{E}_{x,y}|l(h(x), y) - l(h^*(x), y)| - 2L^* \\ &\geq \mathbf{E}_{x,y} C_0 |y(h(x) - h^*(x))| - 2L^* \\ &= C_0 \mathbf{E}_x |h(x) - h^*(x)| - 2L^*. \end{aligned}$$

So we can upperbound $\max_{h \in H_t} \mathbf{E}_x |h(x) - h^*(x)|$ (in terms of L^* and Δ_{t-1}), whereas we want to upperbound the expected value of p_t, which is proportional

to $\mathbf{E}_x \max_{h \in H_t} |h(x) - h^*(x)|$. The ratio between these two quantities is related to a fundamental parameter of the learning problem, a generalization of the *disagreement coefficient* (Hanneke, 2007).

We flesh out this intuition in the remainder of this section. First we describe a broader class of loss functions than those considered above (including 0–1 loss, which is not differentiable); a distance metric on hypotheses, and a generalized disagreement coefficient. We then prove that for this broader class, active learning performs better than passive learning when the generalized disagreement coefficient is small.

5.2. A subclass of loss functions

We give label complexity upper bounds for a class of loss functions that includes 0–1 loss and logistic loss but not hinge loss. Specifically, we require that the loss function has bounded *slope asymmetry*, defined below.

Recall earlier notation: response space Z, classifier space $H = \{h : X \to Z\}$, and loss function $l : Z \times Y \to [0, \infty)$. Henceforth, the label space is $Y = \{-1, +1\}$.

Definition 4. *The* slope asymmetry *of a loss function $l : Z \times Y \to [0, \infty)$ is*

$$K_l = \sup_{z, z' \in Z} \frac{\max_{y \in Y} |l(z, y) - l(z', y)|}{\min_{y \in Y} |l(z, y) - l(z', y)|}.$$

The slope asymmetry is 1 for 0–1 loss, and ∞ for hinge loss. For differentiable loss functions $l(z, y) = \phi(yz)$, it is easily related to bounds on the derivative.

Lemma 5. *Let $l_\phi(z, y) = \phi(zy)$, where ϕ is a differentiable function defined on $Z = (-B, B) \subset \mathbb{R}$. Suppose $C_0 \le |\phi'(z)| \le C_1$ for all $z \in Z$. Then for any $z, z' \in Z$, and any $y \in \{-1, +1\}$,*

$$C_0 |z - z'| \le |l_\phi(z, y) - l_\phi(z', y)| \le C_1 |z - z'|.$$

Thus l_ϕ has slope asymmetry $\le C_1/C_0$.

Proof. By the mean value theorem, there is some $\xi \in Z$ such that $l_\phi(z, y) - l_\phi(z', y) = \phi(yz) - \phi(yz') = \phi'(\xi)(yz - yz')$. Thus $|l_\phi(z, y) - l_\phi(z', y)| = |\phi'(\xi)| \cdot |z - z'|$, and the rest follows from the bounds on ϕ'. □

For instance, this immediately applies to logistic loss.

Corollary 6. *Logistic loss $l(z, y) = \ln(1 + e^{-yz})$, defined on label space $Y = \{-1, +1\}$ and response space $[-B, B]$, has slope asymmetry at most $1 + e^B$.*

5.3. Topologizing the space of classifiers

We introduce a simple distance function on the space of classifiers.

Definition 7. *For any $f, g \in H$ and distribution D define $\rho(f, g) = \mathbf{E}_{x \sim D} \max_y |l(f(x), y) - l(g(x), y)|$. For any $r \ge 0$, let $B(f, r) = \{g \in H : \rho(f, g) \le r\}$.*

Suppose $L^* = \min_{h \in H} L(h)$ is realized at h^*. We know that at time t, the remaining hypotheses have loss at most $L^* + 2\Delta_{t-1}$. Does this mean they are close to h^* in ρ-distance? The ratio between the two can be expressed in terms of the slope asymmetry of the loss.

Lemma 8. *For any distribution D and any loss function with slope asymmetry K_l, we have $\rho(h, h^*) \le K_l(L(h) + L^*)$ for all $h \in H$.*

Proof. For any $h \in H$,

$$\begin{aligned}
\rho(h, h^*) &= \mathbf{E}_x \max_y |l(h(x), y) - l(h^*(x), y)| \\
&\le K_l \, \mathbf{E}_{x,y} |l(h(x), y) - l(h^*(x), y)| \\
&\le K_l \, (\mathbf{E}_{x,y}[l(h(x), y)] + \mathbf{E}_{x,y}[l(h^*(x), y)]) \\
&= K_l \, (L(h) + L(h^*)).
\end{aligned}$$

□

5.4. A generalized disagreement coefficient

When analyzing the A^2 algorithm (Balcan et al., 2006) for active learning under 0–1 loss, (Hanneke, 2007) found that its label complexity could be characterized in terms of what he called the *disagreement coefficient* of the learning problem. We now generalize this notion to arbitrary loss functions.

Definition 9. *The disagreement coefficient is the infimum value of θ such that for all r,*

$$\mathbf{E}_{x \sim D} \sup_{h \in B(h^*, r)} \sup_y |l(h(x), y) - l(h^*(x), y)| \le \theta r.$$

Here is a simple example for linear separators.

Lemma 10. *Suppose H consists of linear classifiers $\{u \in \mathbb{R}^d : \|u\| \le B\}$ and the data distribution D is uniform over the surface of the unit sphere in $\mathbb{R}^d$. Suppose the loss function is $l(z, y) = \phi(yz)$ for differentiable ϕ with $C_0 \le |\phi'| \le C_1$. Then the disagreement coefficient is at most $(2C_1/C_0)\sqrt{d}$.*

Proof. Let h^* be the optimal classifier, and h any other classifier with $\rho(h, h^*) \le r$. Let u^*, u be the corresponding vectors in $\mathbb{R}^d$. Using lemma 5, $r \ge \mathbf{E}_{x \sim D} \sup_y |l(h(x), y) - l(h^*(x), y)| \ge C_0 \, \mathbf{E}_{x \sim D} |h(x) - h^*(x)| = C_0 \, \mathbf{E}_{x \sim D} |(u - u^*) \cdot x| \ge C_0 \, \|u - u^*\| / (2\sqrt{d})$. Thus for any $h \in B(h^*, r)$, we have that the corresponding vectors satisfy $\|u - u^*\| \le 2r\sqrt{d}/C_0$. We can now bound the disagreement coefficient. $\mathbf{E}_{x \sim D} \sup_{h \in B(h^*, r)} \sup_y |l(h(x), y) - l(h^*(x), y)| \le C_1 \, \mathbf{E}_{x \sim D} \sup_{h \in B(h^*, r)} |h(x) - h^*(x)| \le C_1 \, \mathbf{E}_x \sup\{|(u - u^*) \cdot x| : \|u - u^*\| \le 2r\sqrt{d}/C_0\} \le C_1 \cdot 2r\sqrt{d}/C_0$. □

5.5. Upper Bound on Label Complexity

Finally, we give a bound on label complexity for learning problems with bounded disagreement coefficient and loss functions with bounded slope asymmetry.

Theorem 11. *For all learning problems D and hypothesis spaces H, if the loss function has slope asymmetry K_l, and the learning problem has disagreement coefficient θ, then for all $\delta > 0$, with probability at least $1-\delta$ over the choice of data, the expected number of labels requested by IWAL(loss-weighting) during the first T iterations is at most*

$$4\theta \cdot K_l \cdot (L^*T + O(\sqrt{T \ln(|H|T/\delta)})),$$

where L^* is the minimum loss achievable on D by H, and the expectation is over the randomness in the selective sampling.

Proof. Suppose $h^* \in H$ achieves loss L^*. Pick any time t. By theorem 2, $H_t \subset \{h \in H : L(h) \leq L^* + 2\Delta_{t-1}\}$ and by lemma 8, $H_t \subset B(h^*, r)$ for $r = K_l(2L^* + 2\Delta_{t-1})$. Thus, the expected value of p_t (over the choice of x at time t) is at most $\mathbf{E}_{x\sim D} \sup_{f,g\in H_t} \sup_y |l(f(x),y) - l(g(x),y)| \leq 2\,\mathbf{E}_{x\sim D} \sup_{h\in H_t} \sup_y |l(h(x),y) - l(h^*(x),y)| \leq 2\,\mathbf{E}_{x\sim D} \sup_{h\in B(h^*,r)} \sup_y |l(h(x),y) - l(h^*(x),y)| \leq 2\theta r = 4\theta \cdot K_l \cdot (L^* + \Delta_{t-1})$. Summing over $t = 1, \ldots, T$, we get the lemma, using $\sum_{t=1}^{T}(1/\sqrt{t}) = O(\sqrt{T})$. □

6. A lower bound on label complexity

Kääriäinen (2006) showed that for any hypothesis class H and any $\eta > \epsilon > 0$, there is a data distribution such that (a) the optimal error rate achievable by H is η; and (b) any active learner that finds $h \in H$ with error rate at most $\eta + \epsilon$ (with probability greater than $1/2$) must make η^2/ϵ^2 queries. We now strengthen this lower bound to $d\eta^2/\epsilon^2$, where d is the VC dimension of H.

Let's see how this relates to the label complexity rates of the previous section. It is well-known that if a supervised learner sees T examples (for any $T > d/\eta$), its final hypothesis has error at most $\eta + \sqrt{d\eta/T}$ (Devroye et al., 1996) with high probability. Think of this as $\eta + \epsilon$ for $\epsilon = \sqrt{d\eta/T}$. Our lower bound now implies that an active learner must make at least $d\eta^2/\epsilon^2 = \eta T$ queries. This explains the ηT leading term in all the label complexity bounds we have discussed.

Theorem 12. *For any $\eta, \epsilon > 0$ such that $2\epsilon \leq \eta \leq 1/4$, for any input space X and hypothesis class H (of functions mapping X into $Y = \{+1, -1\}$) of VC dimension $1 < d < \infty$, there is a distribution over $X \times Y$ such that (a) the best error rate achievable by H is η; (b) any active learner seeking a classifier of error at most $\eta + \epsilon$ must make $\Omega(d\eta^2/\epsilon^2)$ queries to succeed with probability at least $1/2$.*

Proof. Pick a set of d points $x_o, x_1, x_2, \ldots, x_{d-1}$ shattered by H. Here is a distribution over $X \times Y$: point x_o has probability $1 - \beta$, while each of the remaining x_i has probability $\beta/(d-1)$, where $\beta = 2(\eta + 2\epsilon)$. At x_o, the response is always $y = 1$. At $x_i, i \geq 1$, the response is $y = 1$ with probability $1/2 + \gamma b_i$, where b_i is either $+1$ or -1, and $\gamma = 2\epsilon/\beta = \epsilon/(\eta + 2\epsilon) < 1/4$.

Nature starts by picking $b_1, \ldots, b_{d-1}$ uniformly at random. This defines the target hypothesis h^*: $h^*(x_o) = 1$ and $h^*(x_i) = b_i$. Its error rate is $\beta \cdot (1/2 - \gamma) = \eta$.

Any learner outputs a hypothesis in H and thus implicitly makes guesses at the underlying hidden bits b_i. Unless it correctly determines b_i for at least $3/4$ of the points $x_1, \ldots, x_{d-1}$, the error of its hypothesis will be at least $\eta + (1/4) \cdot \beta \cdot (2\gamma) = \eta + \epsilon$.

Now, suppose the active learner makes at most $c(d-1)/\gamma^2$ queries, where c is a small constant ($c \leq 1/125$ suffices). We'll show that it fails (outputs a hypothesis with error at least $\eta + \epsilon$) with probability at least $1/2$.

We'll say x_i is *heavily queried* if the active learner queries it at least $4c/\gamma^2$ times. At most $1/4$ of the x_i's are heavily queried; without loss of generality, these are $x_1, \ldots, x_k$, for some $k \leq (d-1)/4$. The remaining x_i get so few queries that the learner guesses each corresponding bit b_i with probability less than $2/3$; this can be derived from Slud's lemma (below), which relates the tails of a binomial to that of a normal.

Let F_i denote the event that the learner gets b_i wrong; so $\mathbf{E}F_i \geq 1/3$ for $i > k$. Since $k \leq (d-1)/4$, the probability that the learner fails is given by $\mathbf{P}[F_1 + \cdots + F_{d-1} \geq (d-1)/4] \geq \mathbf{P}[F_{k+1} + \cdots + F_{d-1} \geq (d-1)/4] \geq \mathbf{P}[B \geq (d-1)/4] \geq \mathbf{P}[Z \geq 0] = 1/2$, where B is a binomial$((3/4)(d-1), 1/3)$ random variable, Z is a standard normal, and the last inequality follows from Slud's lemma. Thus the active learner must make at least $c(d-1)/\gamma^2 = \Omega(d\eta^2/\epsilon^2)$ queries to succeed with probability at least $1/2$. □

Lemma 13 (Slud (1977))**.** *Let B be a Binomial (n, p) random variable with $p \leq 1/2$, and let Z be a standard normal. For any $k \in [np, n(1-p)]$, $\mathbf{P}[B \geq k] \geq \mathbf{P}[Z \geq (k - np)/\sqrt{np(1-p)}]$.*

Theorem 12 uses the same example that is used for lower bounds on supervised sample complexity (section 14.4 of (Devroye et al., 1996)), although in that case the lower bound is $d\eta/\epsilon^2$. The bound for active learning is smaller by a factor of η because the ac-

tive learner can avoid making repeated queries to the "heavy" point x_o, whose label is immediately obvious.

7. Implementing IWAL

IWAL(loss-weighting) can be efficiently implemented in the case where H is the class of bounded-length linear separators $\{u \in \mathbb{R}^d : \|u\|^2 \leq B\}$ and the loss function is convex: $l(z, y) = \phi(yz)$ for convex ϕ.

Each iteration of Algorithm 2 involves solving two optimization problems over a restricted hypothesis set

$$H_t = \bigcap_{t'<t} \left\{ h \in H : \frac{1}{t'} \sum_{i=1}^{t'} \frac{Q_i}{p_i} l(h(x_i), y_i) \leq L^*_{t'} + \Delta_{t'} \right\}.$$

Replacing each h by its corresponding vector u, this is $\bigcap_{t'<t}\{u \in \mathbb{R}^d : \frac{1}{t'}\sum_{i=1}^{t'} \frac{Q_i}{p_i}\phi(u \cdot (y_i x_i)) \leq L^*_{t'} + \Delta_{t'}, \|u\|^2 \leq B\}$, an intersection of convex constraints.

The first optimization in Algorithm 2 is $L^*_T = \min_{u \in H_T} \sum_{i=1}^{T} \frac{Q_i}{p_i}\phi(u \cdot (y_i x_i))$, a convex program.

The second optimization is $\max_{u,v \in H_T} \phi(y(u \cdot x)) - \phi(y(v \cdot x))$, $y \in \{+1, -1\}$ (where u, v correspond to functions f, g). If ϕ is nonincreasing (as it is for 0–1, hinge, or logistic loss), then the solution of this problem is $\max\{\phi(A(x)) - \phi(-A(-x)), \phi(A(-x)) - \phi(-A(x))\}$, where $A(x)$ is the solution of a convex program: $A(x) \equiv \min_{u \in H_T} u \cdot x$. The two cases inside the max correspond to the choices $y = 1$ and $y = -1$.

Thus Algorithm 2 can be efficiently implemented for nonincreasing convex loss functions and bounded-length linear separators. In our experiments, we use a simpler implementation. For the first problem (determining L^*_T), we minimize over H rather than H_T; for the second (determining $A(x)$), instead of defining H_T by $T-1$ convex constraints, we simply enforce the last of these constraints (corresponding to time $T-1$). This may lead to an overly conservative choice of p_t, but by theorem 1, the consistency of h_T is assured.

7.1. Experiments

Recent consistent active learning algorithms (Balcan et al., 2006; Dasgupta et al., 2008) have suffered from computational intractability. This section shows that importance weighted active learning is practical.

We implemented IWAL with loss-weighting for linear separators under logistic loss. As outlined above, the algorithm involves two convex optimizations as subroutines. These were coded using log-barrier methods (section 11.2 of (Boyd & Vandenberghe, 2004)). We tried out the algorithm on the MNIST data set of handwritten digits by picking out the 3's and 5's as two classes, and choosing 1000 exemplars of each for training and another 1000 of each for testing. We used PCA to reduce the dimension from 784 to 25. The algorithm uses a generalization bound Δ_t of the form $\sqrt{d/t}$; since this is believed to often be loose in high dimensions, we also tried a more optimistic bound of $1/\sqrt{t}$. In either case, active learning achieved very similar performance (in terms of test error or test logistic loss) to a supervised learner that saw all the labels. The active learner asked for less than 1/3 of the labels.

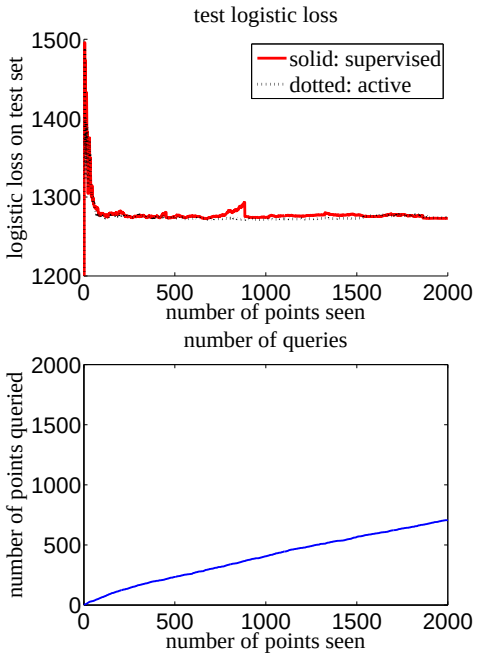

Figure 1. Top: Test logistic loss as number of points seen grows from 0 to 2000. Bottom: Number of queries versus number of points seen.

7.2. Bootstrap instantiation of IWAL

This section reports another practical implementation of IWAL, using a simple bootstrapping scheme: Query an initial set of points, and use the resulting labeled examples to train a set of predictors H. Given a new unlabeled example x, the sampling probability is set to $p_{\min} + (1 - p_{\min})\big[\max_{y;h_i,h_j \in H} L(h_i(x), y) - L(h_j(x), y)\big]$, where $p_{\min}$ is a lower bound on the sampling probability.

We implemented this scheme for binary and multiclass classification loss, using 10 decision trees trained on the initial 1/10th of the training set, setting $p_{\min} = 0.1$. For simplicity, we did't retrain the predictors for each new queried point, i.e., the predictors were trained once on the initial sample. The final predictor is trained on the collected importance-weighted

training set, and tested on the test set. The Costing technique (Zadrozny et al., 2003) was used to remove the importance weights using rejection sampling. (The same technique can be applied to any loss function.) The resulting unweighted classification problem was then solved using a decision tree learner (J48). On the same MNIST dataset as in section 7.1, the scheme performed at least as well as passive learning, using only 65.6% of the labels (Beygelzimer et al., 2008).

8. Conclusion

The IWAL algorithms and analysis presented here remove many reasonable objections to the deployment of active learning. IWAL satisfies the same convergence guarantee as common supervised learning algorithms, it can take advantage of standard algorithms (section 7.2), it can deal with very flexible losses, and in theory and practice it can yield substantial label complexity improvements.

Empirically, in *every* experiment we have tried, IWAL has substantially reduced the label complexity compared to supervised learning, with no sacrifice in performance on the same number of unlabeled examples. Additional experimental results on benchmark datasets are reported in the full version of the paper (Beygelzimer et al., 2008).

Since IWAL explicitly accounts for sample selection bias, we can be sure that these experiments are valid for use in constructing new datasets. This implies another subtle advantage: because the sampling bias is known, it is possible to hypothesize and check the performance of IWAL algorithms on datasets drawn by IWAL. This potential for self-tuning off-policy evaluation is extremely useful when labels are expensive.

References

Abe, N., & Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. *Proceedings of the International Conference on Machine Learning* (pp. 1–9).

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, *68*, 357–367.

Bach, F. (2007). Active learning for misspecified generalized linear models. In *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

Balcan, M.-F., Beygelzimer, A., & Langford, J. (2006). Agnostic active learning. *Proceedings of the International Conference on Machine Learning* (pp. 65–72).

Beygelzimer, A., Dasgupta, S., & Langford, J. (2008). Importance weighted active learning. arXiv:0812.4952v2 [cs.LG].

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

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

Dasgupta, S., & Hsu, D. (2008). Hierarchical sampling for active learning. *Proceedings of the 25th International Conference on Machine learning* (pp. 208–215).

Dasgupta, S., Hsu, D., & Monteleoni, C. (2008). A general agnostic active learning algorithm. In *Advances in neural information processing systems*, vol. 20, 353–360.

Dasgupta, S., Kalai, A. T., & Monteleoni, C. (2005). Analysis of perceptron-based active learning. *Proceedings of the Annual Conference on Learning Theory* (pp. 249–263).

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. *Proceedings of the 24th International Conference on Machine Learning* (pp. 353–360).

Kääriäinen, M. (2006). Active learning in the non-realizable case. *Proceedings of 17th International Conference on Algorithmic Learning Theory* (pp. 63–77).

Langford, J. (2005). Practical prediction theory for classification. *Journal of Machine Learning Research*, *6*, 273–306.

Slud, E. (1977). Distribution inequalities for the binomial law. *Annals of Probability*, *5*, 404–412.

Sugiyama, M. (2006). Active learning for misspecified models. In *Advances in neural information processing systems*, vol. 18, 1305–1312. Cambridge, MA: MIT Press.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Proceedings of the Third IEEE International Conference on Data Mining* (pp. 435–442).