Negative Results for Active Learning with Convex Losses

Steve Hanneke
Department of Statistics,
Carnegie Mellon University

Liu Yang
Machine Learning Department
Carnegie Mellon University

Abstract

We study the problem of active learning with convex loss functions. We prove that even under bounded noise constraints, the minimax rates for proper active learning are often no better than passive learning.

1 Introduction

It is now well established, both empirically and theoretically, that active learning can provide substantial improvements in the convergence rates achievable for classification compared to passive learning for the 0-1 loss function (Dasgupta, 2005; Dasgupta, Hsu, and Monteleoni, 2007; Balcan, Beygelzimer, and Langford, 2006; Hanneke, 2007b,a; Balcan, Hanneke, and Wortman, 2008; Hanneke, 2009b; Tong and Koller, 2001; Beygelzimer, Dasgupta, and Langford, 2009). However, although many positive results on rates of convergence with noisy labels are known, there remains a substantial computational problem in extending these results to practical scenarios, since even passive learning can be computationally difficult in many cases when there is label noise.

In passive learning, one of the primary tricks for avoiding this difficulty is to optimize a surrogate convex loss function, which can be performed computationally efficiently. The hope is that the optimal solution to the surrogate risk will also have small risk under 0-1 loss for many cases (though clearly not always). This is the primary tool that has given rise to such effective passive learning methods as SVM and AdaBoost.

This naturally raises the question of whether this same trick can be employed to create computationally efficient active learning algorithms based on actively optimizing the risk for a convex loss function. The key question is whether active learning will still provide the convergence rate improvements over passive learning, when the loss function is convex rather than discrete. In this work, we explore this issue. Specifically, we find negative results for proper active learning algorithms under a wide variety of convex loss functions, showing that their minimax rates are often no better than the rates achievable for passive learning, even under *bounded* noise conditions, which are known to give favorable results for the 0-1 loss (Hanneke, 2009b; Castro and Nowak, 2006).

The intuition behind these results is that distant points with even small amounts of label noise can dramatically affect the optimal function. This means that the learning algorithm cannot ignore distant points, so that its queries will never become localized to a small region of the instance space (e.g., around the optimal decision boundary). Since the queries never localize, the algorithm cannot improve over passive. Note that this contrasts with 0-1 loss, where distant noisy points are no worse than close noisy points, and an algorithm's queries will often rapidly focus to a region near the optimal function's decision boundary. To make matters worse, the amount to which those distant noisy points affect the optimal function actually depends on the *magnitude* of their noise, so that the learning algorithm essentially needs to estimate the magnitude of the noise in those distant noisy regions in order to properly optimize the risk. Since estimating the magnitude of noise is a task that active learning essentially cannot help with, we are left with a problem active learning is not well suited to solving. Again, this contrasts with the 0-1 loss, where an algorithm can safely ignore any point once it has determined the *sign* of the optimal function's value on that point.

Though negative in nature, we stress that these results should be interpretted carefully. In particular, although we show that active learning is not generally more effective than passive at reducing the risk for convex losses, this does *not* necessarily mean that we cannot design effective active learning algorithms for optimizing the *0-1 loss* based on optimizing a convex *surrogate loss*. Indeed, the "hard" distributions studied below, which give rise to the negative results, are precisely designed so that the risk based on the convex loss is *not* a good approximation to the risk based

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

on the 0-1 loss. This leaves open the question of whether there are efficient active learning algorithms based on optimizing convex surrogate losses, which achieve good rates for the 0-1 loss *when the surrogate risk provides a good approximation*. Furthermore, our results regard the asymptotic dependence on the number of labels, and we leave open the question of significant constant factor improvements (see (Beygelzimer, Dasgupta, and Langford, 2009) for some ideas in this direction).

The remainder of the paper is organized as follows. In Section 2, we introduce basic notation and formalize the learning model. This is followed in Section 3 with a brief survey of convex loss functions and known results for passive learning. In Section 4, we prove our first result, a general negative result for strictly convex losses. In Section 5, we prove a similar negative result for a more general class of convex losses, but only for a particular function class (corresponding to threshold classifiers under 0-1 loss). We conclude in Section 6 with some interpretation and discussion of the results.

2 Definitions and Notation

We consider the problem of optimizing the expectation of a loss function, for a random variable (X, Y) taking values in $\mathcal{X} \times \{-1, +1\}$. Specifically, we suppose there is a distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$, a *loss function* $\ell : \mathbb{R} \to [0, \infty)$, and a *function class* $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$. In this paper, our results will be most interesting for *parametric* classes $\mathcal{F}$ (though our general lower bounds will hold regardless), and we will implicitly assume $\mathcal{F}$ is such throughout the paper. We will primarily be discussing *convex* and *nonincreasing* loss functions: that is, nonincreasing ℓ such that, $\forall x, y \in \mathbb{R}$ and $\alpha \in [0, 1]$, $\ell(\alpha x + (1 - \alpha)y) \leq \alpha\ell(x) + (1 - \alpha)\ell(y)$. For any function $f : \mathcal{X} \to \mathbb{R}$, we define the *risk* $R(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X)Y)]$, the *optimal risk* $R^* = \inf_{f\in\mathcal{F}} R(f)$, and we will refer to the quantity $R(f) - R^*$ as the *excess risk* of f.

In the learning problem, there is a sequence of random variables $(X_1, Y_1), (X_2, Y_2), \ldots$ independent and distributed according to $\mathcal{D}$. In active learning, the algorithm is able to observe the unlabeled sequence $X_1, X_2, \ldots$, then selects an index i_1, receives the label Y_{i_1}, then after observing this label, selects another index i_2, and receives the label Y_{i_2}, etc. Thus, it is only able to observe the Y_i values that it explicitly requests. Suppose that, after each label request, the algorithm produces a function $\hat{f}_n \in \mathcal{F}$ (where n is the number of label requests); note that we are considering *proper* learning algorithms only. We will be interested in the behavior of $\mathbb{E}[R(\hat{f}_n) - R^*]$ as a function of n, where the expectation is over the (X_i, Y_i) sequence and any internal randomness of the algorithm.

The distributions on $Y|X$ we study below are generally referred to as *η-bounded-noise* distributions; they satisfy the property that $\exists f \in \mathcal{F}$ such that, $\forall x \in \mathcal{X}$,

$$\mathbb{P}(Y \neq sign(f(x))|X = x) \leq \eta < 1/2.$$

These conditions are considered favorable toward learning, and in particular active learning can often achieve exponential rates of convergence for the 0-1 loss function under bounded noise distributions (Hanneke, 2009a; Castro and Nowak, 2006).

2.1 A Sampler of Loss Functions

The primary loss function studied in this classification setting is the 0-1 loss: $\ell(x) = \mathbb{1}[x \leq 0]$. Thus, $\ell(yf(x))$ corresponds to testing whether $sign(f(x)) \neq y$, and the risk becomes $\mathbb{P}_{(X,Y)\sim\mathcal{D}}(f(X) \neq Y)$. However, as this loss is not convex, and thus often computationally difficult to optimize, we are often interested in convex *relaxations* of the 0-1 loss, often referred to as a *surrogate* loss function.

As we are interested in losses that are surrogates for 0-1 loss, in this work, we will restrict ourselves to loss functions that are nonincreasing, as is the case for those losses mentioned below; these are often associated with margin-based learning algorithms, since the loss increases as the function value of an incorrect prediction increases, and decreases as the function value of a correct prediction increases.

Perhaps the most well-known such surrogate loss is the *hinge loss*: $\ell(x) = \max\{1 - x, 0\}$. In some sense, this represents the "least convex" loss function that upper bounds the 0-1 loss. It is used extensively in margin-based learning algorithms. Other common convex loss functions are the *exponential loss* $\ell(x) = e^{-x}$, used in AdaBoost, and the logistic loss, $\ell(x) = \log(1 + e^{-x})$, used in logistic regression.

3 Known Results for Passive Learning

In this context, a *passive learning* algorithm is simply any active learning algorithm that requests the labels in the order they appear in the original sequence: that is, $i_1 = 1$, $i_2 = 2$, and generally $i_n = n$.

Convex loss functions have been studied in some depth in the passive learning literature. In particular, (Bartlett, Jordan, and Mcauliffe, 2005) provide several nice results under low noise conditions. Most relevant to our present discussion, they show that for strictly convex losses (with at least a certain "modulus of convexity") and a few additional constraints on $\mathcal{F}$ and ℓ, under bounded noise distributions, there are passive learning algorithms such that $\mathbb{E}[R(\hat{f}_n) - R^*] \leq c/n$, for some constant c depending on ℓ, $\mathcal{F}$, and η. Thus, these results will serve as our baseline for comparison in active learning.

3.1 A Lower Bound for Estimating a Bernoulli Mean

The following well-known result will play a key role in all of our proofs. It lower bounds the minimax risk for estimating the mean of a Bernoulli random variable.

Lemma 1. *For $0 \leq a < b \leq 1$, there exists a constant $c_{a,b} > 0$ such that, for any $n \in \mathbb{N}$, for any estimator $\hat{p}_n : \{0,1\}^n \to [0,1]$, there exists a value $p \in [a,b]$ such that, if $B_1, B_2, \ldots, B_n$ are independent $Bernoulli(p)$ random variables, then*

$$\mathbb{E}[(\hat{p}_n(B_1, B_2, \ldots, B_n) - p)^2] > c_{a,b}/n.$$

4 General Results for Strictly Convex Losses

Let ℓ be twice differentiable, with $\forall x \in \mathbb{R}, \ell(x) > 0, \ell'(x) < 0, \ell''(x) > 0$, and with all three of these everywhere continuous. Furthermore, let $\mathcal{F}$ be a function class with the property that, for some $x_0 \in \mathcal{X}$, $[1/2, 1] \subseteq \{f(x_0) : f \in \mathcal{F}\}$.

Theorem 1. *For any ℓ and $\mathcal{F}$ satisfying the above conditions, there exists a distribution on $\mathcal{X}$, a noise bound $\eta \in [0, 1/2)$, and a constant $c > 0$ such that, for any active learning algorithm $\hat{f}_n$ and any $n \in \mathbb{N}$, there is an η-bounded-noise label distribution for which*

$$\mathbb{E}[R(\hat{f}_n) - R^*] > c/n.$$

Note that, since passive learning can achieve a rate c'/n for certain strictly convex losses under the stated conditions (e.g., (Bartlett, Jordan, and Mcauliffe, 2005)), this represents a negative result.

Proof of Theorem 1. The proof is by reduction from estimating the mean of a Bernoulli random variable.

Define the distribution on $\mathcal{X}$ so that $\mathbb{P}(\{x_0\}) = 1$. In particular, all of the active learning algorithm's queries must be at x_0. Then parametrize the label distribution by $\nu = \mathbb{P}(Y = -1 | X = x_0) \in [0, 1]$. In particular, the equation $\nu\ell'(-y) = (1-\nu)\ell'(y)$, derived by setting the derivative of the conditional risk given x_0 equal to zero, defines a continuous bijection $\phi : (0,1) \to \mathbb{R}$ between ν values in $(0,1)$ and y values in $\mathbb{R}$. Thus, there exists a range of ν values $[\nu_1, \nu_{1/2}] = \phi^{-1}([1/2, 1])$, for which the risk minimizer f^* has $f^*(x_0) \in [1/2, 1]$. Furthermore, $0 < \nu_1 < \nu_{1/2} < 1/2$, and we can take $\eta = \nu_{1/2}$. Now, given any active learning algorithm $\hat{f}_n$, define an estimator $\hat{\nu}_n$ for ν as follows. If $\hat{f}_n(x_0) \in [1/2, 1]$, define $\tilde{f}_n(x_0) = \hat{f}_n(x_0)$; otherwise, define $\tilde{f}_n(x_0) = \operatorname{argmin}_{y \in \{1/2, 1\}} |y - \hat{f}_n(x_0)|$. Then define $\hat{\nu}_n = \phi^{-1}(\tilde{f}_n(x_0))$. Finally, let $\nu^* \in [\nu_1, \nu_{1/2}]$ be the value for which $\mathbb{E}[(\hat{\nu}_n - \nu^*)^2] > c_{\nu_1, \nu_{1/2}}/n$, guaranteed to exist by Lemma 1, and define

$$\mathbb{P}(Y = -1 | X = x_0) = \nu^*.$$

Letting f^* be such that $f^*(x_0) = \phi(\nu^*)$, we have

$$R(\hat{f}_n) - R(f^*) \geq R(\tilde{f}_n) - R(f^*).$$

Noting that compactness of $[-1, 1]$ implies ℓ'' is uniformly continuous on $[-1, 1]$, let $m = \inf_{y \in [-1,1]} \ell''(y) > 0$ and $M = \sup_{y \in [-1,1]} \ell''(y) < \infty$; in particular, this also means ℓ is strongly convex on $[-1, 1]$. Let $\tilde{y} = \tilde{f}_n(x_0)$ and $y^* = f^*(x_0)$. Then

$$\begin{aligned}
&R(\tilde{f}_n) - R(f^*) \\
&= (1-\nu^*)(\ell(\tilde{y}) - \ell(y^*)) + \nu^*(\ell(-\tilde{y}) - \ell(-y^*)) \\
&\geq ((1-\nu^*)\ell'(y^*) - \nu^*\ell'(-y^*))(\tilde{y} - y^*) \\
&+ (1-\nu^*)\frac{m}{2}(\tilde{y} - y^*)^2 \\
&= (1-\nu^*)\frac{m}{2}(\tilde{y} - y^*)^2.
\end{aligned}$$

Also, for any $a < b$, there is some $c \in [a, b]$ such that $\ell'(b) - \ell'(a) = \ell''(c)(b - a) \leq M(b - a)$. Therefore,

$$\begin{aligned}
&|\hat{\nu}_n - \nu^*| \cdot |\ell'(1/2)| \\
&\leq |\hat{\nu}_n - \nu^*||\ell'(-\tilde{y}) + \ell'(\tilde{y})| \\
&= \nu^*|\ell'(-y^*) - \ell'(-\tilde{y})| + (1-\nu^*)|\ell'(\tilde{y}) - \ell'(y^*)| \\
&\leq \nu^* M|\tilde{y} - y^*| + (1-\nu^*)M|\tilde{y} - y^*| = M|\tilde{y} - y^*|.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}[R(\hat{f}_n) - R(f^*)] \\
&\geq (1-\nu^*)\frac{m}{2}\mathbb{E}[(\tilde{y} - y^*)^2] \\
&\geq (1-\nu^*)\frac{m\ell'(1/2)^2}{2M^2}\mathbb{E}[(\hat{\nu}_n - \nu^*)^2] \\
&> (1-\nu^*)\frac{m\ell'(1/2)^2}{2M^2}c_{\nu_1,\nu_2}/n.
\end{aligned}$$

□

It is straightforward to relax the stated conditions considerably; for instance, we can allow a few discontinuities in ℓ', or require the existence of any arbitrary interval $[y_0, y_1]$ of $f(x_0)$ values (rather than the arbitrary $[1/2, 1]$). Also, it is easy to generalize this result to essentially any reasonable distribution for most natural function classes. Note that the distribution presently described is essentially trivial to learn under the 0-1 loss (with exponential rates).

5 Constant-Slope One-Dimensional Linear Functions with General Convex Losses

The results of the previous section were proven for strictly convex losses. However, the intuitive idea seems to hold for any convex loss function, and the results of this section, though in some ways less general in that they are proven for only one particular function class $\mathcal{F}$, are also in some ways more general in that they hold for loss functions that might not be strictly convex, nor be everywhere differentiable, nor be strictly positive nor strictly decreasing. The example also serves to highlight the intuition that distant noisy points are typically the cause of poor performance for active learning with convex losses.

Specifically, in this section we will consider any nonincreasing continuous convex ℓ that satisfies $\ell(0) > 0$, ℓ is differentiable at 0 and has $\ell'(0) < 0$, is differentiable at all but at most a finite number of points, and has $\lim_{x\to\infty} \ell(x) = 0$. Furthermore, we will denote by $z > 0$ some point for which ℓ is differentiable at z and $\ell'(z) > \ell'(0)/2$.

We will also restrict ourselves to the specific class of functions

$$\mathcal{F} = \{f_t(x) = x - t : t \in \mathbb{R}\},$$

which in some sense corresponds to the "thresholds" function class commonly studied for the 0-1 loss (and known to allow exponential rates for bounded noise in that case).

Theorem 2. *Suppose $\mathcal{F}$ and ℓ are as described above. There exists a distribution on $\mathcal{X}$ and a constant $c > 0$ such that, for any active learning algorithm $\hat{f}_n$ and any $n \in \mathbb{N}$, there is a $1/4$-bounded-noise label distribution for which*

$$\mathbb{E}[R(\hat{f}_n) - R^*] > c/n.$$

Proof of Theorem 2. Define the marginal distribution on $\mathbb{R}$ by a density that is $1/(4z)$ in $[0, 4z]$, and 0 elsewhere. For the conditional distribution of Y given X, parameterize it by a value $\nu \in [0, 1/4]$, and let $\mathbb{P}(Y = +1|X = x) = 1$ for $x \le 2z$, and $\mathbb{P}(Y = +1|X = x) = 1 - \nu$ for $x > 2z$. For each such distribution based on a given ν, let $R_\nu(\cdot)$ represent the corresponding risk functional, and R^*_ν the corresponding optimal risk. For each $f_t \in \mathcal{F}$, let ν_t denote the value of ν for which $R_\nu(f_t) = R^*_\nu$. Also, there exists some range $[a, b]$ with $0 \le a < b \le 1/4$ in which $\operatorname{argmin}_t R_\nu(f_t) \in (0, z)$ for all $\nu \in [a, b]$. For a given learning algorithm $\hat{f}_n$, define $\hat{\nu}_n = \nu_t$ such that $\hat{f}_n = f_t$.

Now construct a reduction from the Bernoulli mean estimation problem, by constructing the conditional $Y|X$ distribution as follows. Given a $p \in [0, 1/4]$, and a sequence $B_1, B_2, \ldots, B_n$ of independent $Bernoulli(p)$ random variables, on the i^{th} time the active learning algorithm requests a label in $(2z, \infty)$, return -1 if $B_i = 1$ and otherwise return $+1$. For any query in $(-\infty, 2z]$, simply return $+1$. This corresponds to taking $\nu = p$ in the conditional distribution. Now let ν^* be the value of p for which $\mathbb{E}[(\hat{\nu}_n - \nu^*)^2] > c_{0,1/4}/n$ when we use ν^* in the conditional of Y given X; such a ν^* is guaranteed to exist by Lemma 1.

Next, we lower bound the excess risk as a function of this distance. Note that $\frac{\partial R_{\nu^*}(f_{t^*})}{\partial t^*} = 0$, where t^* is such that $R(f_{t^*}) = R^*$. Now for $t \in (0, z)$ for which ℓ is differentiable at $2z - t$ and $4z - t$,

$$\begin{aligned}
\frac{\partial R_{\nu^*}(f_t)}{\partial t} &\propto -[\nu^*\ell(t-4z) + \nu^*\ell(2z-t) \\
&+ (1-\nu^*)\ell(4z-t) - (\ell(-t) + \nu^*\ell(t-2z))] \\
&= -[\nu^*[(\ell(t-4z) - \ell(t^*-4z)) \\
&+ (\ell(2z-t) - \ell(2z-t^*)) - (\ell(4z-t) - \ell(4z-t^*)) \\
&- (\ell(t-2z) - \ell(t^*-2z))] \\
&+ [(\ell(4z-t) - \ell(4z-t^*)) - (\ell(-t) - \ell(-t^*))]].
\end{aligned}$$

To simplify things, suppose $t - t^* \le 0$; the other case is proven analogously. By basic convexity inequalities, the magnitude of the above is at least as big as

$$\begin{aligned}
&\nu^*(t-t^*)[\ell'(t^*-4z) - \ell'(t^*-2z) \\
&+ \ell'(4z-t^*) - \ell'(2z-t^*)] \\
&+ (t-t^*)\left[-\ell'(4z-t^*) + \ell'(-t^*)\right]. \qquad (1)
\end{aligned}$$

Since

$$\ell'(t^*-2z) - \ell'(t^*-4z) > \ell'(4z-t^*) - \ell'(2z-t^*) > 0,$$

we have

$$\ell'(t^*-4z) - \ell'(t^*-2z) + \ell'(4z-t^*) - \ell'(2z-t^*) < 0.$$

Also,

$$\ell'(4z-t^*) - \ell'(-t^*) > -\ell'(0)/2 = |\ell'(0)|/2.$$

Thus,

$$\left|\frac{\partial R_{\nu^*}(f_t)}{\partial t}\right| \ge c'|\ell'(0)| \cdot |t - t^*|,$$

for some constant c'.

A similar argument shows this remains true for $t - t^* > 0$. Specifically, we replace (1) by an analogous bound, using the reverse inequality $\ell(y) - \ell(x) \le \ell'(y)(y - x)$ (which effectively keeps the terms that involve t, rather than those involving t^* as in (1)).

Thus,

$$\begin{aligned}
R_{\nu^*}(\hat{f}_n) - R^* &\ge \left|\int_0^{\hat{t}_n - t^*} \frac{\partial R_{\nu^*}(f_{x+t^*})}{\partial x} \mathrm{d}x\right| \\
&\ge c''|\hat{t}_n - t^*|^2.
\end{aligned}$$

Additionally noting that

$$\begin{aligned}
0 = \frac{\partial R_{\nu^*}(f_t)}{\partial t}|_{t=t^*} - \frac{\partial R_{\hat{\nu}_n}(f_t)}{\partial t}|_{t=\hat{t}_n} \\
\ge c_3[(\hat{t}_n - t^*) - \\
(\hat{\nu}_n - \nu^*)(\ell(2z-\hat{t}_n) - \ell(4z-\hat{t}_n) + \ell(\hat{t}_n-4z) - \ell(\hat{t}_n-2z))],
\end{aligned}$$

and similarly for $\frac{\partial R_{\hat{\nu}_n}(f_t)}{\partial t}|_{t=\hat{t}_n} - \frac{\partial R_{\nu^*}(f_t)}{\partial t}|_{t=t^*}$, we generally have $|\hat{t}_n - t^*| \ge c_4|\hat{\nu}_n - \nu^*|$. Thus, $\mathbb{E}[R(\hat{f}_n) - R^*] \ge c_5\mathbb{E}[(\hat{\nu}_n - \nu^*)^2] > c_6/n$. $\square$

6 Discussion

It should clearly be possible to relax many of these conditions on the losses and generalize to other types of distributions. However, the point here is simply to highlight the fact that we generally should not expect improvements for convex losses due to the dragging effect of distant noisy points on the solution.

That said, as mentioned earlier, the negative results proven above should be interpretted carefully. In particular, although we have shown that active learning is not generally more effective at reducing risk based on convex losses, this does not necessarily mean that an active learning algorithm designed to optimize a convex loss will not achieve improved rates for the 0-1 loss. Nonetheless, it seems that great care would be needed in designing active learning algorithms for convex surrogate losses; in particular, unlike existing algorithms that directly optimize the 0-1 loss (Balcan, Beygelzimer, and Langford, 2006; Dasgupta, Hsu, and Monteleoni, 2007; Beygelzimer, Dasgupta, and Langford, 2009), an algorithm based on a convex surrogate loss should not necessarily be designed to optimize its *worst-case* performance under the *surrogate* loss. For such an algorithm, even when the noise distribution is such that the surrogate approximates the 0-1 loss, the mere *possibility* of distant noisy points may prevent the algorithm from focusing its queries, thus preventing improvements over passive learning; that is, algorithms that optimize for the worst-case scenario for the surrogate will tend to *search* for noisy regions. Rather, to be an effective algorithm for the 0-1 loss, we should optimize the algorithm's performance for only the noise distributions under which the solutions to the surrogate loss are also good for the 0-1 loss. Since typically these are the only scenarios we are interested in anyway, the use of surrogate losses may yet be a viable possibility for designing efficient and effective active learning algorithms. The details of designing and analyzing such methods are left for future work.

References

Balcan, M.-F., Beygelzimer, A., and Langford, J. (2006). Agnostic active learning. In *Proc. of the 23rd International Conference on Machine Learning*.

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

Bartlett, P., Jordan, M. I., and Mcauliffe, J. (2005). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, **101**, 138–156.

Beygelzimer, A., Dasgupta, S., and Langford, J. (2009). Importance weighted active learning.

Castro, R. and Nowak, R. (2006). Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*.

Dasgupta, S., Hsu, D., and Monteleoni, C. (2007). A general agnostic active learning algorithm. Technical Report CS2007-0898, Department of Computer Science and Engineering, University of California, San Diego.

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

Hanneke, S. (2007b). Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Conference on Learning Theory*.

Hanneke, S. (2009a). Adaptive rates of convergence in active learning. In *Proceedings of the 22nd Conference on Learning Theory*.

Hanneke, S. (2009b). *Theoretical Foundations of Active Learning*. Ph.D. thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University.

Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, **2**.