Journal of Machine Learning Research 13 (2012) 255-279 Submitted 10/10; Revised 7/11; Published 2/12

Active Learning via Perfect Selective Classification

Ran El-Yaniv RANI@CS.TECHNION.AC.IL
Yair Wiener WYAIR@TX.TECHNION.AC.IL
Computer Science Department
Technion – Israel Institute of Technology
Haifa 32000, Israel

Editor: Yoav Freund

Abstract

We discover a strong relation between two known learning models: stream-based active learning and perfect selective classification (an extreme case of 'classification with a reject option'). For these models, restricted to the realizable case, we show a reduction of active learning to selective classification that preserves fast rates. Applying this reduction to recent results for selective classification, we derive exponential target-independent label complexity speedup for actively learning general (non-homogeneous) linear classifiers when the data distribution is an arbitrary high dimensional mixture of Gaussians. Finally, we study the relation between the proposed technique and existing label complexity measures, including teaching dimension and disagreement coefficient.

Keywords: classification with a reject option, perfect classification, selective classification, active learning, selective sampling, disagreement coefficient, teaching dimension, exploration vs. exploitation

1. Introduction and Related Work

Active learning is an intriguing learning model that provides the learning algorithm with some control over the learning process, potentially leading to significantly faster learning. In recent years it has been gaining considerable recognition as a vital technique for efficiently implementing inductive learning in many industrial applications where abundance of unlabeled data exists, and/or in cases where labeling costs are high. In this paper we expose a strong relation between active learning and *selective classification*, another known alternative learning model (Chow, 1970; El-Yaniv and Wiener, 2010).

Focusing on binary classification in realizable settings we consider standard *stream-based active learning*, which is also referred to as *online selective sampling* (Atlas et al., 1990; Cohn et al., 1994). In this model the learner is given an error objective ε and then sequentially receives unlabeled examples. At each step, after observing an unlabeled example x, the learner decides whether or not to request the label of x. The learner should terminate the learning process and output a binary classifier whose true error is guaranteed to be at most ε with high probability. The penalty incurred by the learner is the number of label requests made and this number is called the *label complexity*. A label complexity bound of $O(d\log(d/\varepsilon))$ for actively learning ε-good classifier from a concept class with VC-dimension d, provides an exponential speedup in terms of $1/\varepsilon$ relative to standard (passive) supervised learning where the sample complexity is typically $O(d/\varepsilon)$.

The study of (stream-based, realizable) active learning is paved with very interesting theoretical results. Initially, only a few cases were known where active learning provides significant advan-

tage over passive learning. Perhaps the most favorable result was an exponential label complexity speedup for learning homogeneous linear classifiers where the (linearly separable) data is uniformly distributed over the unit sphere. This result was manifested by various authors using various analysis techniques, for a number of strategies that can all be viewed in hindsight as approximations or variations of the "CAL algorithm" of Cohn et al. (1994). Among these studies, the earlier theoretical results (Seung et al., 1992; Freund et al., 1993, 1997; Fine et al., 2002; Gilad-Bachrach, 2007) considered Bayesian settings and studied the speedup obtained by the Query by Committee (QBC) algorithm. The more recent results provided PAC style analyses (Dasgupta et al., 2009; Hanneke, 2007a, 2009).

Lack of positive results for other non-toy problems, as well as various additional negative results that were discovered, led some researchers to believe that active learning is not necessarily advantageous in general. Among the striking negative results is Dasgupta's negative example for actively learning general (non-homogeneous) linear classifiers (even in two dimensions) under the uniform distribution over the sphere (Dasgupta, 2005).

A number of recent innovative papers proposed alternative models for active learning. Balcan et al. (2008) introduced a subtle modification of the traditional label complexity definition, which opened up avenues for new positive results. According to their new definition of "non-verifiable" label complexity, the active learner is not required to know when to stop the learning process with a guaranteed ε-good classifier. Their main result, under this definition, is that active learning is asymptotically better than passive learning in the sense that only $o(1/\varepsilon)$ labels are required for actively learning an ε-good classifier from a concept class that has a finite VC-dimension. Another result they accomplished is an exponential label complexity speedup for (non-verifiable) active learning of non-homogeneous linear classifiers under the uniform distribution over the the unit sphere.

Based on Hanneke's characterization of active learning in terms of the "disagreement coefficient" (Hanneke, 2007a), Friedman (2009) recently extended the Balcan et al. results and proved that a target-dependent exponential speedup can be asymptotically achieved for a wide range of "smooth" learning problems (in particular, the hypothesis class, the instance space and the distribution should all be expressible by smooth functions). He proved that under such smoothness conditions, for any target hypothesis h^*, Hanneke's disagreement coefficient is bounded above in terms of a constant $c(h^*)$ that depends on the unknown target hypothesis h^* (and is independent of δ and ε). The resulting label complexity is $O\left(c(h^*)\, d \,\mathrm{polylog}(d/\varepsilon)\right)$ (Hanneke, 2011b). This is a very general result but the *target-dependent* constant involved in this bound is only guaranteed to be finite.

With this impressive progress in the case of target-dependent bounds for active learning, the current state of affairs in the *target-independent* bounds for active learning arena leaves much to be desired. To date the most advanced result in this model, which was already essentially established by Seung et al. and Freund et al. more than fifteen years ago (Seung et al., 1992; Freund et al., 1993, 1997), is still a target-independent exponential speed up bound for homogeneous linear classifiers under the uniform distribution over the sphere.

The other learning model we contemplate that will be shown to have strong ties to active learning, is *selective classification*, which is mainly known in the literature as 'classification with a reject option.' This old-timer model, that was already introduced more than fifty years ago (Chow, 1957, 1970), extends standard supervised learning by allowing the classifier to opt out from predictions in cases where it is not confident. The incentive is to increase classification reliability over instances that are not rejected by the classifier. Thus, using selective classification one can potentially achieve

a lower error rate using the same labeling "budget." The main quantities that characterize a selective classifier are its (true) error and coverage rate (or its complement, the rejection rate).

There is already substantial volume of research publications on selective classification, that kept emerging through the years. The main theme in many of these publications is the implementation of certain reject mechanisms for specific learning algorithms like support vector machines and neural networks. Among the few theoretical studies on selective classification, there are various excess risk bounds for ERM learning (Herbei and Wegkamp, 2006; Bartlett and Wegkamp, 2008; Wegkamp, 2007), and certain coverage/risk guarantees for selective ensemble methods (Freund et al., 2004). In a recent work (El-Yaniv and Wiener, 2010) the trade-off between error and coverage was examined and in particular, a new extreme case of selective learning was introduced. In this extreme case, termed here "perfect selective classification," the classifier is given m labeled examples and is required to instantly output a classifier whose true error is perfectly zero with certainty. This is of course potentially doable only if the classifier rejects a sufficient portion of the instance space. A non-trivial result for perfect selective classification is a high probability lower bound on the classifier coverage (or equivalently, an upper bound on its rejection rate). Such bounds have recently been presented in El-Yaniv and Wiener (2010).

In Section 3 we present a reduction of active learning to perfect selective classification that preserves "fast rates." This reduction enables the luxury of analyzing *dynamic* active learning problems as *static* problems. Relying on a recent result on perfect selective classification from El-Yaniv and Wiener (2010), in Section 4 we then apply our reduction and conclude that general (non-homogeneous) linear classifiers are actively learnable at exponential (in $1/\varepsilon$) label complexity rate when the data distribution is an arbitrary unknown finite mixture of high dimensional Gaussians. While we obtain exponential label complexity speedup in $1/\varepsilon$, we incur exponential slowdown in d^2, where d is the problem dimension. Nevertheless, in Section 5 we prove a lower bound of $\Omega((\log m)^{(d-1)/2}(1+o(1))$ on the label complexity, when considering the class of unrestricted linear classifiers under a Gaussian distribution. Thus, an exponential slowdown in d is unavoidable in such settings.

Finally, in Section 6 we relate the proposed technique to other complexity measures for active learning. Proving and using a relation to the *teaching dimension* (Goldman and Kearns, 1995) we show, by relying on a known bound for the teaching dimension, that perfect selective classification with meaningful coverage can be achieved for the case of axis-aligned rectangles under a product distribution. We then focus on Hanneke's *disagreement coefficient* and show that the coverage of perfect selective classification can be bounded below using the disagreement coefficient. Conversely, we show that the disagreement coefficient can be bounded above using any coverage bound for perfect selective classification. Consequently, the results here imply that the disagreement coefficient can be sufficiently bounded to ensure fast active learning for the case of linear classifiers under a mixture of Gaussians.

2. Active Learning and Perfect Selective Classification

In *binary classification* the goal is to learn an accurate *binary classifier*, $h : \mathcal{X} \to \{\pm 1\}$, from a finite labeled training sample. Here $\mathcal{X}$ is some instance space and the standard assumption is that the training sample, $S_m = \{(x_i, y_i)\}_{i=1}^m$, containing m labeled examples, is drawn i.i.d. from some unknown distribution $P(X,Y)$ defined over $\mathcal{X} \times \{\pm 1\}$. The classifier h is chosen from some hypothesis class $\mathcal{H}$. In this paper we focus on the *realizable setting* whereby labels are defined by

some unknown *target hypothesis* $h^* \in \mathcal{H}$. Thus, the underlying distribution reduces to $P(X)$. The performance of a classifier h is quantified by its true zero-one *error*, $R(h) \triangleq \Pr\{h(X) \neq h^*(X)\}$. A positive result for a classification problem $(\mathcal{H}, P)$ is a learning algorithm that given an error target ε and a confidence parameter δ can output, based on S_m, an hypothesis h whose error $R(h) \leq \varepsilon$, with probability of at least $1-\delta$. A bound $B(\varepsilon, \delta)$ on the size m of labeled training sample sufficient for achieving this is called the *sample complexity* of the learning algorithm. A classical result is that any consistent learning algorithm has sample complexity of $O(\frac{1}{\varepsilon}(d\log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})))$, where d is the VC-dimension of $\mathcal{H}$ (see, e.g., Anthony and Bartlett, 1999).

2.1 Active Learning

We consider the following standard active learning model. In this model the learner sequentially observes unlabeled instances, $x_1, x_2, \ldots$, that are sampled i.i.d. from $P(X)$. After receiving each x_i, the learning algorithm decides whether or not to request its label $h^*(x_i)$, where $h^* \in \mathcal{H}$ is an unknown target hypothesis. Before the start of the game the algorithm is provided with some desired error rate ε and confidence level δ. We say that the learning algorithm *actively learned* the problem instance $(\mathcal{H}, P)$ if at some point it can terminate this process, after observing m instances and requesting k labels, and output an hypothesis $h \in \mathcal{H}$ whose error $R(h) \leq \varepsilon$, with probability of at least $1-\delta$. The quality of the algorithm is quantified by the number k of requested labels, which is called the *label complexity*. A positive result for a learning problem $(\mathcal{H}, P)$ is a learning algorithm that can actively learn this problem for any given ε and δ, and for every h^*, with label complexity bounded above by $L(\varepsilon, \delta, h^*)$. If there is a label complexity bound that is $O(polylog(1/\varepsilon))$ we say that the problem is *actively learnable at exponential rate*.

2.2 Selective Classification

Following the formulation in El-Yaniv and Wiener (2010) the goal in selective classification is to learn a pair of functions (h, g) from a labeled training sample S_m (as defined above for passive learning). The pair (h, g), which is called a *selective classifier*, consists of a binary classifier $h \in \mathcal{H}$, and a *selection function*, $g : \mathcal{X} \to \{0, 1\}$, which qualifies the classifier h as follows. For any sample $x \in \mathcal{X}$, the output of the selective classifier is $(h, g)(x) \triangleq h(x)$ iff $g(x) = 1$, and $(h, g)(x) \triangleq \mathit{abstain}$ iff $g(x) = 0$. Thus, the function g is a filter that determines a sub-domain of $\mathcal{X}$ over which the selective classifier will abstain from classifications. A selective classifier is thus characterized by its *coverage*, $\Phi(h, g) \triangleq \mathbf{E}_P\{g(x)\}$, which is the P-weighted volume of the sub-domain of $\mathcal{X}$ that is not filtered out, and its *error*, $R(h, g) = \mathbf{E}\{\mathbb{I}(h(X) \neq h^*(X)) \cdot g(X)\}/\Phi(h, g)$, which is the zero-one loss restricted to the covered sub-domain. Note that this is a "smooth" generalization of passive learning and, in particular, $R(h, g)$ reduces to $R(h)$ (standard classification) if $g(x) \equiv 1$. We expect to see a trade-off between $R(h, g)$ and $\Phi(h, g)$ in the sense that smaller error should be obtained by compromising the coverage. A major issue in selective classification is how to optimally control this trade-off. In this paper we are concerned with an extreme case of this trade-off whereby (h, g) is required to achieve a perfect score of *zero error with certainty*. This extreme learning objective is termed *perfect learning* in El-Yaniv and Wiener (2010). Thus, for a *perfect selective classifier* (h, g) we always have $R(h, g) = 0$, and its quality is determined by its guaranteed coverage. A positive result for (perfect) selective classification problem $(\mathcal{H}, P)$ is a learning algorithm that uses a labeled training sample S_m (as in passive learning) to output a perfect selective classifier (h, g) for which $\Phi(h, g) \geq B_\Phi(\mathcal{H}, \delta, m)$ with probability of at least $1-\delta$, for any given δ. The bound

$B_\Phi = B_\Phi(\mathcal{H}, \delta, m)$ is called a *coverage bound* (or *coverage rate*) and its complement, $1 - B_\Phi$, is called a *rejection bound* (or *rate*). A coverage rate $B_\Phi = 1 - O(\frac{polylog(m)}{m})$ (and the corresponding $1 - B_\Phi$ rejection rate) are qualified as *fast*.

2.3 The CAL Algorithm and the Consistent Selective Strategy (CSS)

The major players in active learning and in perfect selective classification are the CAL algorithm and the consistent selective strategy (CSS), respectively. To define them we need the following definitions.

Definition 1 (Version space, Mitchell, 1977) *Given an hypothesis class $\mathcal{H}$ and a training sample S_m, the* version space *$VS_{\mathcal{H},S_m}$ is the set of all hypotheses in $\mathcal{H}$ that classify S_m correctly.*

Definition 2 (Disagreement set, Hanneke, 2007a; El-Yaniv and Wiener, 2010) *Let $\mathcal{G} \subset \mathcal{H}$. The* disagreement set *w.r.t. $\mathcal{G}$ is defined as*

$$DIS(\mathcal{G}) \triangleq \{x \in \mathcal{X} : \exists h_1, h_2 \in \mathcal{G} \quad s.t. \quad h_1(x) \neq h_2(x)\}.$$

The agreement set *w.r.t. $\mathcal{G}$ is $AGR(\mathcal{G}) \triangleq \mathcal{X} \setminus DIS(\mathcal{G})$.*

The main strategy for active learning in the realizable setting (Cohn et al., 1994) is to request labels only for instances belonging to the disagreement set and output any (consistent) hypothesis belonging to the version space. This strategy is often called the *CAL algorithm*. A related strategy for perfect selective classification was proposed in El-Yaniv and Wiener (2010) and termed *consistent selective strategy (CSS)*. Given a training set S_m, CSS takes the classifier h to be any hypothesis in $VS_{\mathcal{H},S_m}$ (i.e., a consistent learner), and takes a selection function g that equals one for all points in the agreement set with respect to $VS_{\mathcal{H},S_m}$, and zero otherwise.

3. From Coverage Bound to Label Complexity Bound

In this section we present a reduction from stream-based active learning to perfect selective classification. Particularly, we show that if there exists for $\mathcal{H}$ a perfect selective classifier with a fast rejection rate of $O(polylog(m)/m)$, then the CAL algorithm will actively learn $\mathcal{H}$ with exponential label complexity rate of $O(polylog(1/\varepsilon))$.

Lemma 3 *Let $S_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sequence of m labeled samples drawn i.i.d. from an unknown distribution $P(X)$ and let $S_i = \{(x_1, y_1), \ldots, (x_i, y_i)\}$ be the i-prefix of S_m. Then, with probability of at least $1 - \delta$ over random choices of S_m, the following bound holds simultaneously for all $i = 1, \ldots, m-1$,*

$$\Pr\left\{x_{i+1} \in DIS(VS_{\mathcal{H},S_i}) | S_i\right\} \leq 1 - B_\Phi\left(\mathcal{H}, \frac{\delta}{\log_2(m)}, 2^{\lfloor \log_2(i) \rfloor}\right),$$

where $B_\Phi(\mathcal{H}, \delta, m)$ is a coverage bound for perfect selective classification with respect to hypothesis class $\mathcal{H}$, confidence δ and sample size m .

Proof For $j = 1, \ldots, m$, abbreviate $DIS_j \triangleq DIS(VS_{\mathcal{H},S_j})$ and $AGR_j \triangleq AGR(VS_{\mathcal{H},S_j})$. By definition, $DIS_j = \mathcal{X} \setminus AGR_j$. By the definitions of a coverage bound and agreement/disagreement sets, with probability of at least $1-\delta$ over random choices of S_j

$$B_\Phi(\mathcal{H}, \delta, j) \leq \Pr\{x \in AGR_j | S_j\} = \Pr\{x \notin DIS_j | S_j\} = 1 - \Pr\{x \in DIS_j | S_j\}.$$

Applying the union bound we conclude that the following inequality holds simultaneously with high probability for $t = 0, \ldots, \lfloor \log_2(m) \rfloor - 1$,

$$\Pr\{x_{2^t+1} \in DIS_{2^t} | S_{2^t}\} \leq 1 - B_\Phi\left(\mathcal{H}, \frac{\delta}{\log_2(m)}, 2^t\right). \tag{1}$$

For all $j \leq i$, $S_j \subseteq S_i$, so $DIS_i \subseteq DIS_j$. Therefore, since the samples in S_m are all drawn i.i.d., for any $j \leq i$,

$$\Pr\left\{x_{i+1} \in DIS_i | S_i\right\} \leq \Pr\left\{x_{i+1} \in DIS_j | S_j\right\} = \Pr\left\{x_{j+1} \in DIS_j | S_j\right\}.$$

The proof is complete by setting $j = 2^{\lfloor \log_2(i) \rfloor} \leq i$, and applying inequality (1). ■

Lemma 4 (Bernstein's inequality Hoeffding, 1963) *Let $X_1, \ldots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for all positive t,*

$$\Pr\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{t^2/2}{\sum \mathbf{E}\left\{X_j^2\right\} + Mt/3}\right).$$

Lemma 5 *Let Z_i, $i = 1, \ldots, m$, be independent Bernoulli random variables with success probabilities p_i. Then, for any $0 < \delta < 1$, with probability of at least $1-\delta$,*

$$\sum_{i=1}^{m} (Z_i - \mathbf{E}\{Z_i\}) \leq \sqrt{2 \ln \frac{1}{\delta} \sum p_i} + \frac{2}{3} \ln \frac{1}{\delta}.$$

Proof Define $W_i \triangleq Z_i - \mathbf{E}\{Z_i\} = Z_i - p_i$. Clearly,

$$\mathbf{E}\{W_i\} = 0, \quad |W_i| \leq 1, \quad \mathbf{E}\{W_i^2\} = p_i(1-p_i).$$

Applying Bernstein's inequality (Lemma 4) on the W_i,

$$\begin{aligned} \Pr\left\{\sum_{i=1}^{n} W_i > t\right\} &\leq \exp\left(-\frac{t^2/2}{\sum \mathbf{E}\left[W_j^2\right] + t/3}\right) = \exp\left(-\frac{t^2/2}{\sum p_i(1-p_i) + t/3}\right) \\ &\leq \exp\left(-\frac{t^2/2}{\sum p_i + t/3}\right). \end{aligned}$$

Equating the right-hand side to δ and solving for t, we have

$$\frac{t^2/2}{\sum p_i + t/3} = \ln \frac{1}{\delta} \quad \Longleftrightarrow \quad t^2 - t \cdot \frac{2}{3} \ln \frac{1}{\delta} - 2 \ln \frac{1}{\delta} \sum p_i = 0,$$

and the positive solution of this quadratic equation is

$$t = \frac{1}{3}\ln\frac{1}{\delta} + \sqrt{\frac{1}{9}\ln^2\frac{1}{\delta} + 2\ln\frac{1}{\delta}\sum p_i} < \frac{2}{3}\ln\frac{1}{\delta} + \sqrt{2\ln\frac{1}{\delta}\sum p_i}.$$

■

Lemma 6 *Let $Z_1, Z_2, \ldots, Z_m$ be a high order Markov sequence of dependent binary random variables defined in the same probability space. Let $X_1, X_2, \ldots, X_m$ be a sequence of independent random variables such that,*

$$\Pr\{Z_i = 1 | Z_{i-1}, \ldots, Z_1, X_{i-1}, \ldots, X_1\} = \Pr\{Z_i = 1 | X_{i-1}, \ldots, X_1\}.$$

Define $P_1 \triangleq \Pr\{Z_1 = 1\}$, and for $i = 2, \ldots, m$,

$$P_i \triangleq \Pr\{Z_i = 1 | X_{i-1}, \ldots, X_1\}.$$

Let $b_1, b_2 \ldots b_m$ be given constants independent of $X_1, X_2, \ldots, X_m$.[1] *Assume that $P_i \leq b_i$ simultaneously for all i with probability of at least $1 - \delta/2$, $\delta \in (0,1)$. Then, with probability of at least $1 - \delta$,*

$$\sum_{i=1}^{m} Z_i \leq \sum_{i=1}^{m} b_i + \sqrt{2\ln\frac{2}{\delta}\sum b_i} + \frac{2}{3}\ln\frac{2}{\delta}.$$

We proceed with a direct proof of Lemma 6. An alternative proof of this lemma, using supermartingales, appears in Appendix B.

Proof For $i = 1, \ldots, m$, let W_i be binary random variables satisfying

$$\begin{aligned}
&\Pr\{W_i = 1 | Z_i = 1, X_{i-1}, \ldots, X_1\} \triangleq \frac{b_i + \mathbb{I}(P_i \leq b_i) \cdot (P_i - b_i)}{P_i}, \\
&\Pr\{W_i = 1 | Z_i = 0, X_{i-1}, \ldots, X_1\} \triangleq \max\left\{\frac{b_i - P_i}{1 - P_i}, 0\right\}, \\
&\Pr\{W_i = 1 | W_{i-1}, \ldots, W_1, X_{i-1}, \ldots, X_1\} = \Pr\{W_i = 1 | X_{i-1}, \ldots, X_1\}.
\end{aligned}$$

We notice that

$$\begin{aligned}
\Pr\{W_i = 1 | X_{i-1}, \ldots, X_1\} &= \Pr\{W_i = 1, Z_i = 1 | X_{i-1}, \ldots, X_1\} \\
&+ \Pr\{W_i = 1, Z_i = 0 | X_{i-1}, \ldots, X_1\} \\
&= \Pr\{W_i = 1 | Z_i = 1, X_{i-1}, \ldots, X_1\} \Pr\{Z_i = 1 | X_{i-1}, \ldots, X_1\} \\
&+ \Pr\{W_i = 1 | Z_i = 0, X_{i-1}, \ldots, X_1\} \Pr\{Z_i = 0 | X_{i-1}, \ldots, X_1\} \\
&= \begin{cases} P_i + \frac{b_i - P_i}{1 - P_i}(1 - P_i) = b_i, & P_i \leq b_i; \\ \frac{b_i}{P_i} \cdot P_i + 0 = b_i, & \text{else.} \end{cases}
\end{aligned}$$

Hence the distribution of each W_i is independent of $X_{i-1}, \ldots, X_1$, and the W_i are independent Bernoulli random variables with success probabilities b_i. By construction if $P_i \leq b_i$ then

$$\Pr\{W_i = 1 | Z_i = 1\} = \int_X \Pr\{W_i = 1 | Z_i = 1, X_{i-1}, \ldots, X_1\} = 1.$$

1. Precisely we require that each of the b_i were selected before X_i are chosen

By assumption $P_i \leq b_i$ for all i simultaneously with probability of at least $1-\delta/2$. Therefore, $Z_i \leq W_i$ simultaneously with probability of at least $1-\delta/2$. We now apply Lemma 5 on the W_i. The proof is then completed using the union bound. ■

Theorem 7 *Let S_m be a sequence of m unlabeled samples drawn i.i.d. from an unknown distribution P. Then with probability of at least $1-\delta$ over choices of S_m, the number of label requests k by the CAL algorithm is bounded by*

$$k \leq \Psi(\mathcal{H},\delta,m) + \sqrt{2\ln\frac{2}{\delta}\Psi(\mathcal{H},\delta,m)} + \frac{2}{3}\ln\frac{2}{\delta},$$

where

$$\Psi(\mathcal{H},\delta,m) \triangleq \sum_{i=1}^{m}\left(1 - B_\Phi\left(\mathcal{H}, \frac{\delta}{2\log_2(m)}, 2^{\lfloor \log_2(i)\rfloor}\right)\right)$$

and $B_\Phi(\mathcal{H},\delta,m)$ is a coverage bound for perfect selective classification with respect to hypothesis class $\mathcal{H}$, confidence δ and sample size m .

Proof According to CAL, the label of sample x_i will be requested iff $x_i \in DIS(VS_{\mathcal{H},S_{i-1}})$. For $i = 1,\ldots,m$, let Z_i be binary random variables such that $Z_i \triangleq 1$ iff CAL requests a label for sample x_i. Applying Lemma 3 we get that for all $i = 2,\ldots,m$, with probability of at least $1-\delta/2$

$$\Pr\{Z_i = 1|S_{i-1}\} \quad = \quad \Pr\left\{x_i \in DIS(VS_{\mathcal{H},S_{i-1}})|S_{i-1}\right\} \leq 1 - B_\Phi\left(\mathcal{H}, \frac{\delta}{2\log_2(m)}, 2^{\lfloor \log_2(i-1)\rfloor}\right).$$

For $i = 1$, $B_\Phi(\mathcal{H},\delta,1) = 0$ and the above inequality trivially holds. An application of Lemma 6 on the variables Z_i completes the proof. ■

Theorem 7 states an upper bound on the label complexity expressed in terms of m, the size of the sample provided to CAL. This upper bound is very convenient for directly analyzing the active learning speedup relative to supervised learning. A standard label complexity upper bound, which depends on $1/\varepsilon$, can be extracted using the following simple observation.

Lemma 8 (Hanneke, 2009; Anthony and Bartlett, 1999) *Let S_m be a sequence of m unlabeled samples drawn i.i.d. from an unknown distribution P. Let $\mathcal{H}$ be a hypothesis class whose finite VC dimension is d, and let ε and δ be given. If*

$$m \geq \frac{4}{\varepsilon}\left(d\ln\frac{12}{\varepsilon} + \ln\frac{2}{\delta}\right),$$

then, with probability of at least $1-\delta$, CAL will output a classifier whose true error is at most ε.

Proof Hanneke (2009) observed that since CAL requests a label whenever there is a disagreement in the version space, it is guaranteed that after processing m examples, CAL will output a classifier that is consistent with all the m examples introduced to it. Therefore, CAL is a consistent learner. A classical result (Anthony and Bartlett, 1999, Theorem 4.8) is that any consistent learner will achieve, with probability of at least $1-\delta$, a true error not exceeding ε after observing at most $\frac{4}{\varepsilon}\left(d\ln\frac{12}{\varepsilon} + \ln\frac{2}{\delta}\right)$ labeled examples. ■

Theorem 9 *Let $\mathcal{H}$ be a hypothesis class whose finite VC dimension is d. If the rejection rate of CSS (see definition in Section 2.3) is $O\left(\frac{polylog\left(\frac{m}{\delta}\right)}{m}\right)$, then $(\mathcal{H},P)$ is actively learnable with exponential label complexity speedup.*

Proof Plugging this rejection rate into Ψ (defined in Theorem 7) we have,

$$\Psi(\mathcal{H},\delta,m) \triangleq \sum_{i=1}^{m}\left(1-B_{\Phi}(\mathcal{H},\frac{\delta}{\log_2(m)},2^{\lfloor\log_2(i)\rfloor})\right) = \sum_{i=1}^{m} O\left(\frac{polylog\left(\frac{i\log(m)}{\delta}\right)}{i}\right).$$

Applying Lemma 41 we get

$$\Psi(\mathcal{H},\delta,m) = O\left(polylog\left(\frac{m\log(m)}{\delta}\right)\right).$$

By Theorem 7, $k = O\left(polylog\left(\frac{m}{\delta}\right)\right)$, and an application of Lemma 8 concludes the proof. ■

4. Label Complexity Bounding Technique and Its Applications

In this section we present a novel technique for deriving target-independent label complexity bounds for active learning. The technique combines the reduction of Theorem 7 and a general data-dependent coverage bound for selective classification from El-Yaniv and Wiener (2010). For some learning problems it is a straightforward technical exercise, involving VC-dimension calculations, to arrive with exponential label complexity bounds. We show a few applications of this technique resulting in both reproductions of known label complexity exponential rates as well as a new one. The following definitions (El-Yaniv and Wiener, 2010) are required for introducing the technique.

Definition 10 (Version space compression set) *For any hypothesis class $\mathcal{H}$, let S_m be a labeled sample of m points inducing a version space $VS_{\mathcal{H},S_m}$. The* version space compression set, *$S' \subseteq S_m$, is a smallest subset of S_m satisfying $VS_{\mathcal{H},S_m} = VS_{\mathcal{H},S'}$. The (unique) number $\hat{n} = \hat{n}(\mathcal{H},S_m) = |S'|$ is called the version space compression set* size.

Remark 11 *Our "version space compression set" is precisely Hanneke's "minimum specifying set" (Hanneke, 2007b) for f on U with respect to V, where,*

$$f = h^*, \quad U = S_m, \quad V = \mathcal{H}[S_m] \quad \textit{(see Definition 23)}.$$

Definition 12 (Characterizing hypothesis) *For any subset of hypotheses $\mathcal{G} \subseteq \mathcal{H}$, the* characterizing hypothesis *of $\mathcal{G}$, denoted $f_{\mathcal{G}}(x)$, is a binary hypothesis over $\mathcal{X}$ (not restricted to $\mathcal{H}$) obtaining positive values over the agreement set $AGR(\mathcal{G})$ (Definition 2), and zero otherwise.*

Definition 13 (Order-n characterizing set) *For each n, let Σ_n be the set of all possible labeled samples of size n (all n-subsets, each with all 2^n possible labelings). The order-n* characterizing set *of $\mathcal{H}$, denoted $\mathcal{F}_n$, is the set of all characterizing hypotheses $f_{\mathcal{G}}(x)$, where $\mathcal{G} \subseteq \mathcal{H}$ is a version space induced by some member of Σ_n.*

Definition 14 (Characterizing set complexity) *Let $\mathcal{F}_n$ be the order-n characterizing set of $\mathcal{H}$. The order-n* characterizing set complexity *of $\mathcal{H}$, denoted $\gamma(\mathcal{H},n)$, is the VC-dimension of $\mathcal{F}_n$.*

The following theorem, credited to (El-Yaniv and Wiener, 2010, Theorem 21), is a powerful data-dependent coverage bound for perfect selective learning, expressed in terms of the version space compression set size and the characterizing set complexity.

Theorem 15 (Data-dependent coverage guarantee) *For any m, let $a_1,a_2,\ldots,a_m \in \mathbb{R}$ be given, such that $a_i \geq 0$ and $\sum_{i=1}^m a_i \leq 1$. Let (h,g) be perfect selective classifier (CSS, see Section 2.3). Then, $R(h,g)=0$, and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(h,g) \geq 1 - \frac{2}{m}\left[\gamma(\mathcal{H},\hat{n})\ln_+\left(\frac{2em}{\gamma(\mathcal{H},\hat{n})}\right) + \ln\frac{2}{a_{\hat{n}}\delta}\right],$$

where $\hat{n}$ is the size of the version space compression set and $\gamma(\mathcal{H},\hat{n})$ is the order-$\hat{n}$ characterizing set complexity of $\mathcal{H}$.

Given an hypothesis class $\mathcal{H}$, our recipe to deriving active learning label complexity bounds for $\mathcal{H}$ is: (i) calculate both $\hat{n}$ and $\gamma(\mathcal{H},\hat{n})$; (ii) apply Theorem 15, obtaining a bound B_Φ for the coverage; (iii) plug B_Φ in Theorem 7 to get a label complexity bound expressed as a summation; (iv) Apply Lemma 41 to obtain a label complexity bound in a closed form.

4.1 Examples

In the following example we derive a label complexity bound for the concept class of thresholds (linear separators in $\mathbb{R}$). Although this is a toy example (for which an exponential rate is well known) it does exemplify the technique, and in many other cases the application of the technique is not much harder. Let $\mathcal{H}$ be the class of thresholds. We first show that the corresponding version space compression set size $\hat{n} \leq 2$. Assume w.l.o.g. that $h^*(x) \triangleq \mathbb{I}(x > w)$ for some $w \in (0,1)$. Let $x_- \triangleq \max\{x_i \in S_m | y_i = -1\}$ and $x_+ \triangleq \min(x_i \in S_m | y_i = +1)$. At least one of x_- or x_+ exist. Let $S'_m = \{(x_-,-1),(x_+,+1)\}$. Then $VS_{\mathcal{H},S_m} = VS_{\mathcal{H},S'_m}$, and $\hat{n} = |S'_m| \leq 2$. Now, $\gamma(\mathcal{H},2)=2$, because the order-2 characterizing set of $\mathcal{H}$ is the class of intervals in $\mathbb{R}$ whose VC-dimension is 2. Plugging these numbers in Theorem 15, and using the assignment $a_1 = a_2 = 1/2$,

$$B_\Phi(\mathcal{H},\delta,m) = 1 - \frac{2}{m}\left[2\ln(em) + \ln\frac{4}{\delta}\right] = 1 - O\left(\frac{\ln(m/\delta)}{m}\right).$$

Next we plug B_Φ in Theorem 7 obtaining a raw label complexity

$$\Psi(\mathcal{H},\delta,m) = \sum_{i=1}^m \left(1 - B_\Phi\left(\mathcal{H}, \frac{\delta}{2\log_2(m)}, 2^{\lfloor \log_2(i)\rfloor}\right)\right) = \sum_{i=1}^m O\left(\frac{\ln(\log_2(m)\cdot i/\delta)}{i}\right).$$

Finally, by applying Lemma 41, with $a=1$ and $b = \log_2 m/\delta$, we conclude that

$$\Psi(\mathcal{H},\delta,m) = O\left(\ln^2\left(\frac{m}{\delta}\right)\right).$$

Thus, $\mathcal{H}$ is actively learnable with exponential speedup, and this result applies to any distribution. In Table 1 we summarize the $\hat{n}$ and $\gamma(\mathcal{H},\hat{n})$ values we calculated for four other hypothesis classes. The

Hypothesis class	Distribution	$\hat{n}$	$\gamma(\mathcal{H},\hat{n})$
Linear separators in $\mathbb{R}$	any	2	2
Intervals in $\mathbb{R}$	any (target-dependent)[2]	4	4
Linear separators in $\mathbb{R}^2$	any distribution on the unit circle (target-dependent)[2]	4	4
Linear separators in $\mathbb{R}^d$	mixture of Gaussians	$O\left((\log m)^{d-1}/\delta\right)$	$O\left(\hat{n}^{d/2+1}\right)$
Balanced axis-aligned rectangles in $\mathbb{R}^d$	product distribution	$O(\log(dm/\delta))$	$O(d\hat{n}\log\hat{n})$

Table 1: The $\hat{n}$ and γ of various hypothesis spaces achieving exponential rates.

last two cases are fully analyzed in Sections 4.2 and 6.1, respectively. For the other classes, where γ and $\hat{n}$ are constants, it is clear (Theorem 15) that exponential rates are obtained. We emphasize that the bounds for these two classes are target-dependent as they require that S_m include at least one sample from each class.

4.2 Linear Separators in $\mathbb{R}^d$ Under Mixture of Gaussians

In this section we state and prove our main example, an exponential label complexity bound for linear classifiers in $\mathbb{R}^d$.

Theorem 16 *Let $\mathcal{H}$ be the class of all linear binary classifiers in $\mathbb{R}^d$, and let the underlying distribution be any mixture of a fixed number of Gaussians in $\mathbb{R}^d$. Then, with probability of at least $1-\delta$ over choices of S_m, the number of label requests k by CAL is bounded by*

$$k = O\left(\frac{(\log m)^{d^2+1}}{\delta^{(d+3)/2}}\right).$$

Therefore by Lemma 8 we get $k = O\left(poly(1/\delta)\cdot polylog(1/\varepsilon)\right)$.

Proof The following is a coverage bound for linear classifiers in d dimensions that holds in our setting with probability of at least $1-\delta$ (El-Yaniv and Wiener, 2010, Corollary 33),[3]

$$\Phi(h,g) \geq 1 - O\left(\frac{(\log m)^{d^2}}{m}\cdot\frac{1}{\delta^{(d+3)/2}}\right). \tag{2}$$

2. Target-dependent with at least one sample in each class.
3. This bound uses the fact that for linear classifiers in d dimensions $\hat{n} = O\left((\log m)^{d-1}/\delta\right)$ (El-Yaniv and Wiener, 2010, Lemma 32), and that $\gamma(\mathcal{H},\hat{n}) = O\left(\hat{n}^{d/2+1}\right)$ (El-Yaniv and Wiener, 2010, Lemma 27).

Plugging this bound in Theorem 7 we obtain,

$$\begin{aligned}\Psi(\mathcal{H},\delta,m) &= \sum_{i=1}^{m}\left(1-B_{\Phi}\left(\mathcal{H},\frac{\delta}{2\log_2(m)},2^{\lfloor\log_2(i)\rfloor}\right)\right)\\ &= \sum_{i=1}^{m} O\left(\frac{(\log i)^{d^2}}{i}\cdot\left(\frac{\log_2(m)}{\delta}\right)^{\frac{d+3}{2}}\right)\\ &= O\left(\left(\frac{\log_2(m)}{\delta}\right)^{\frac{d+3}{2}}\cdot\sum_{i=1}^{m}\frac{(\log(i))^{d^2}}{i}\right).\end{aligned}$$

Finally, an application of Lemma 41 with $a=d^2$ and $b=1$ completes the proof. ■

5. Lower Bound on Label Complexity

In the previous section we have derived an upper bound on the label complexity of CAL for various classifiers and distributions. In the case of linear classifiers in $\mathbb{R}^d$ we have shown an exponential speed up in terms of $1/\varepsilon$ but also an exponential slow down in terms of the dimension d. In passive learning there is a linear dependency in the dimension while in our case (active learning using CAL) there is an exponential one. Is it an artifact of our bounding technique or a fundamental phenomenon?

To answer this question we derive an asymptotic lower bound on the label complexity. We show that the exponential dependency in d is unavoidable (at least asymptotically) for every bounding technique when considering linear classifier even under a single Gaussian (isotropic) distribution. The argument is obtained by the observation that CAL has to request a label to any point on the convex hull of a sample S_m. The bound is obtained using known results from probabilistic geometry, which bound the first two moments of the number of vertices of a random polytope under the Gaussian distribution.

Definition 17 (Gaussian polytope) *Let $X_1,\ldots,X_m$ be i.i.d. random points in $\mathbb{R}^d$ with common standard normal distribution (with zero mean and covariance matrix $\frac{1}{2}I_d$). A Gaussian polytope P_m is the convex hull of these random points.*

Denote by $f_k(P_m)$ the number of k-faces in the Gaussian polytope P_m. Note that $f_0(P_m)$ is the number of vertices in P_m. The following two Theorems asymptotically bound the average and variance of $f_k(P_m)$.

Theorem 18 (Hug et al., 2004, Theorem 1.1) *Let $X_1,\ldots,X_m$ be i.i.d. random points in $\mathbb{R}^d$ with common standard normal distribution. Then*

$$\mathbf{E}\,f_k(P_m) = c_{(k,d)}\,(\log m)^{\frac{d-1}{2}}\cdot(1+o(1))$$

as $m\to\infty$, where $c_{(k,d)}$ is a constant depending only on k and d.

Theorem 19 (Hug and Reitzner, 2005, Theorem 1.1) *Let $X_1, \ldots, X_m$ be i.i.d. random points in $\mathbb{R}^d$ with common standard normal distribution. Then there exists a positive constant c_d, depending only on the dimension, such that*

$$Var(f_k(P_m)) \leq c_d (\log m)^{\frac{d-1}{2}}$$

for all $k \in \{0, \ldots, d-1\}$.

We can now use Chebyshev's inequality to lower bound the number of vertices in P_m ($f_0(P_m)$) with high probability.

Theorem 20 *Let $X_1, \ldots, X_m$ be i.i.d. random points in $\mathbb{R}^d$ with common standard normal distribution and $\delta > 0$ be given. Then with probability of at least $1-\delta$,*

$$f_0(P_m) \geq \left(c_d (\log m)^{\frac{d-1}{2}} - \frac{\tilde{c}_d}{\sqrt{\delta}} (\log m)^{\frac{d-1}{4}} \right) \cdot (1 + o(1))$$

as $m \to \infty$, where c_d and $\tilde{c}_d$ are constants depending only on d.

Proof Using Chebyshev's inequality (in the second inequality), as well as Theorem 19 we get

$$\begin{aligned} \Pr(f_0(P_m) > \mathbf{E} f_0(P_m) - t) &= 1 - \Pr(f_0(P_m) \leq \mathbf{E} f_0(P_m) - t) \\ &\geq 1 - \Pr(|f_0(P_m) - \mathbf{E} f_0(P_m)| \geq t) \\ &\geq 1 - \frac{\mathrm{Var}(f_0(P_m))}{t^2} \geq 1 - \frac{c_d}{t^2} (\log m)^{\frac{d-1}{2}}. \end{aligned}$$

Equating the RHS to $1-\delta$ and solving for t we get

$$t = \sqrt{c_d \frac{(\log m)^{\frac{d-1}{2}}}{\delta}}.$$

Applying Theorem 18 completes the proof. ■

Theorem 21 (Lower bound) *Let $\mathcal{H}$ be the class of linear binary classifiers in $\mathbb{R}^d$, and let the underlying distribution be standard normal distribution in $\mathbb{R}^d$. Then there exists a target hypothesis such that, with probability of at least $1-\delta$ over choices of S_m, the number of label requests k by CAL is bounded by*

$$k \geq \frac{c_d}{2} (\log m)^{\frac{d-1}{2}} \cdot (1 + o(1)).$$

as $m \to \infty$, where c_d is a constant depending only on d.

Proof Let us look at the Gaussian polytope P_m induced by the random sample S_m. As long as all labels requested by CAL have the same value (the case of minuscule minority class) we note that every vertex of P_m falls in the region of disagreement with respect to any subset of S_m that do not include that specific vertex. Therefore, CAL will request label at least for each vertex of P_m. For sufficiently large m, in particular,

$$\log m \geq \left(\frac{2\tilde{c}_d}{c_d \sqrt{\delta}} \right)^{\frac{4}{d-1}},$$

we conclude the proof by applying Theorem 20. ■

6. Relation to Existing Label Complexity Measures

A number of complexity measures to quantify the speedup in active learning have been proposed. In this section we show interesting relations between our techniques and two well known measures, namely the teaching dimension (Goldman and Kearns, 1995) and the disagreement coefficient (Hanneke, 2009).

Considering first the teaching dimension, we prove in Lemma 26 that the version space compression set size is bounded above, with high probability, by the extended teaching dimension growth function (introduced by Hanneke, 2007b). Consequently, it follows that perfect selective classification with meaningful coverage can be achieved for the case of axis-aligned rectangles under a product distribution.

We then focus on Hanneke's disagreement coefficient and show in Theorem 34 that the coverage of CSS can be bounded below using the disagreement coefficient. Conversely, in Corollary 39 we show that the disagreement coefficient can be bounded above using any coverage bound for CSS. Consequently, the results here imply that the disagreement coefficient, $\theta(\varepsilon)$ grows slowly with $1/\varepsilon$ for the case of linear classifiers under a mixture of Gaussians.

6.1 Teaching Dimension

The teaching dimension is a label complexity measure proposed by Goldman and Kearns (1995). The dimension of the hypothesis class $\mathcal{H}$ is the minimum number of examples required to present to any consistent learner in order to uniquely identify any hypothesis in the class.

We now define the following variation of the extended teaching dimension (Hegedüs, 1995) due to Hanneke. Throughout we use the notation $h_1(S) = h_2(S)$ to denote the fact that the two hypotheses agree on the classification of all instances in S.

Definition 22 (Extended Teaching Dimension, Hegedüs, 1995; Hanneke, 2007b) *Let $V \subseteq \mathcal{H}$, $m \geq 0$, $U \in \mathcal{X}^m$,*

$$\forall f \in \mathcal{H}, \quad XTD(f,V,U) = \inf\{t \,|\, \exists R \subseteq U : |\{h \in V : h(R) = f(R)\}| \leq 1 \wedge |R| \leq t\}.$$

Definition 23 (Hanneke, 2007b) *For $V \subseteq \mathcal{H}$, $V[S_m]$ denotes any subset of V such that*

$$\forall h \in V, \quad |\{h' \in V[S_m] : h'(S_m) = h(S_m)\}| = 1.$$

Claim 24 *Let S_m be a sample of size m, $\mathcal{H}$ an hypothesis class, and $\hat{n} = n(\mathcal{H}, S_m)$, the version space compression set size. Then,*

$$XTD(h^*, \mathcal{H}[S_m], S_m) = \hat{n}.$$

Proof Let $S_{\hat{n}} \subseteq S_m$ be a version space compression set. Assume, by contradiction, that there exist two hypotheses $h_1, h_2 \in \mathcal{H}[S_m]$, each of which agrees on the given classifications of all examples in $S_{\hat{n}}$. Therefore, $h_1, h_2 \in VS_{\mathcal{H},S_{\hat{n}}}$, and by the definition of version space compression set, we get $h_1, h_2 \in VS_{\mathcal{H},S_m}$. Hence,

$$|\{h \in \mathcal{H}[S_m] : h(S_m) = h^*(S_m)\}| \geq 2,$$

which contradicts definition 23. Therefore,

$$|\{h \in \mathcal{H}[S_m] : h(S_{\hat{n}}) = h^*(S_{\hat{n}})\}| \leq 1,$$

and

$$XTD(h^*, \mathcal{H}[S_m], S_m) \leq |S_{\hat{n}}| = \hat{n}.$$

Let $R \subset S_m$ be any subset of size $|R| < \hat{n}$. Consequently, $VS_{\mathcal{H},S_m} \subset VS_{\mathcal{H},R}$, and there exist hypothesis, $h' \in VS_{\mathcal{H},R}$, that agrees with all labeled examples in R, but disagrees with at least one example in S_m. Thus,

$$h'(S_m) \neq h^*(S_m),$$

and according to definition 23, there exist hypotheses $h_1, h_2 \in \mathcal{H}[S_m]$ such that $h_1(S_m) = h'(S_m) \neq h^*(S_m) = h_2(S_m)$. But $h_1(R) = h_2(R) = h^*(R)$, so

$$|\{h \in V[S_m] : h(R) = h^*(R)\}| \geq 2.$$

It follows that $XTD(h^*, \mathcal{H}[S_m], S_m) \geq \hat{n}$. ■

Definition 25 (XTD Growth Function, Hanneke, 2007b) *For $m \geq 0$, $V \subseteq \mathcal{H}$, $\delta \in [0,1]$,*

$$XTD(V, P, m, \delta) = \inf\left\{t | \forall h \in \mathcal{H}, Pr\left\{XTD(h, V[S_m], S_m) > t\right\} \leq \delta\right\}.$$

Lemma 26 *Let $\mathcal{H}$ be an hypothesis class, P an unknown distribution, and $\delta > 0$. Then, with probability of at least $1-\delta$,*

$$\hat{n} \leq XTD(\mathcal{H}, P, m, \delta).$$

Proof According to Definition 25, with probability of at least $1-\delta$,

$$XTD(h^*, \mathcal{H}[S_m], S_m) \leq XTD(\mathcal{H}, P, m, \delta).$$

Applying Claim 24 completes the proof. ■

Lemma 27 (Balanced Axis-Aligned Rectangles, Hanneke, 2007b, Lemma 4) *If P is a product distribution on $\mathbb{R}^d$ with continuous CDF, and $\mathcal{H}$ is the set of axis-aligned rectangles such that $\forall h \in \mathcal{H}$, $Pr_{X\sim P}\{h(X) = +1\} \geq \lambda$, then,*

$$XTD(\mathcal{H}, P, m, \delta) \leq O\left(\frac{d^2}{\lambda} \log \frac{dm}{\delta}\right).$$

Lemma 28 Blumer et al., 1989, Lemma 3.2.3 *Let $\mathcal{F}$ be a binary hypothesis class of finite VC dimension $d \geq 1$. For all $k \geq 1$, define the k-fold union,*

$$\mathcal{F}_{k\cup} \triangleq \left\{\cup_{i=1}^{k} f_i : f_i \in \mathcal{F}, 1 \leq i \leq k\right\}.$$

Then, for all $k \geq 1$,

$$VC(\mathcal{F}_{k\cup}) \leq 2dk \log_2(3k).$$

Lemma 29 (order-n characterizing set complexity) *Let $\mathcal{H}$ be the class of axis-aligned rectangles in $\mathbb{R}^d$. Then,*

$$\gamma(\mathcal{H}, n) \leq O(dn \log n).$$

Proof Let $S_n = S_k^- \cup S_{n-k}^+$ be a sample of size n composed of k negative examples, $\{x_1, x_2, \ldots x_k\}$, and $n-k$ positive ones. Let $\mathcal{H}$ be the class of axis-aligned rectangles. We define,

$$\forall 1 \leq i \leq k, \qquad R_i \triangleq S_{n-k}^+ \cup \{(x_i, -1)\}.$$

Notice that $VS_{\mathcal{H},R_i}$ includes all axis aligned rectangles that classify all samples in S^+ as positive, and x_i as negative. Therefore, the agreement region of $VS_{\mathcal{H},R_i}$ is composed of two components as depicted in Figure 1. The first component is the smallest rectangle that bounds the positive samples, and the second is an unbounded convex polytope defined by up to d hyperplanes intersecting at x_i. Let AGR_i be the agreement region of $VS_{\mathcal{H},R_i}$ and AGR the agreement region of $VS_{\mathcal{H},S_n}$. Clearly, $R_i \subseteq S_n$, so $VS_{\mathcal{H},S_n} \subseteq VS_{\mathcal{H},R_i}$, and $AGR_i \subseteq AGR$, and it follows that

$$\bigcup_{i=1}^{k} AGR_i \subseteq AGR.$$

Assume, by contradiction, that $x \in AGR$ but $x \notin \bigcup_{i=1}^{k} AGR_i$. Therefore, for any $1 \leq i \leq k$, there exist two hypotheses $h_1^{(i)}, h_2^{(i)} \in VS_{\mathcal{H},R_i}$, such that, $h_1^{(i)}(x) \neq h_2^{(i)}(x)$. Assume, without loss of generality, that $h_1^{(i)}(x) = 1$. We define

$$h_1 \triangleq \bigwedge_{i=1}^{k} h_1^{(i)} \quad \text{and} \quad h_2 \triangleq \bigwedge_{i=1}^{k} h_2^{(i)},$$

meaning that h_1 classifies a sample as positive if and only if all hypotheses $h_1^{(i)}$ classify it as positive. Noting that the intersection of axis-aligned rectangles is itself an axis-aligned rectangle, we know that $h_1, h_2 \in \mathcal{H}$. Moreover, for any x_i we have, $h_1^{(i)}(x_i) = h_2^{(i)}(x_i) = -1$, so also $h_1(x_i) = h_2(x_i) = -1$, and $h_1, h_2 \in VS_{\mathcal{H},S_n}$. But $h_1(x) \neq h_2(x)$. Contradiction. Therefore,

$$AGR = \bigcup_{i=1}^{k} AGR_i.$$

It is well known that the VC dimension of a hyper-rectangle in $\mathbb{R}^d$ is $2d$. The VC dimension of AGR_i is bounded by the VC dimension of the union of two hyper-rectangles in $\mathbb{R}^d$. Furthermore, the VC dimension of AGR is bounded by the VC dimension of the union of all AGR_i. Applying Lemma 28 twice we get,

$$VCdim\{AGR\} \leq 42dk\log_2(3k) \leq 42dn\log_2(3n).$$

If $k = 0$ then the entire sample is positive and the region of agreement is an hyper-rectangle. Therefore, $VCdim\{AGR\} = 2d$. If $k = n$ then the entire sample is negative and the region of agreement is the points of the samples themselves. Hence, $VCdim\{AGR\} = n$. Overall we get that in all cases,

$$VCdim\{AGR\} \leq 42dn\log_2(3n) = O(dn\log n).$$

■

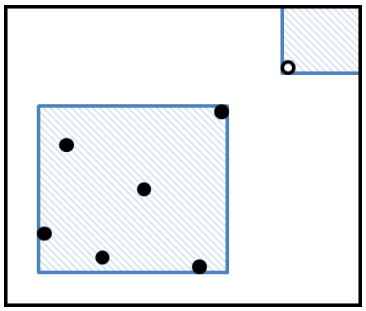

Figure 1: Agreement region of $VS_{\mathcal{H},R_i}$.

Corollary 30 (Balanced Axis-Aligned Rectangles) *Under the same conditions of Lemma 27, the class of balanced axis-aligned rectangles in $\mathbb{R}^d$ can be perfectly selectively learned with fast coverage rate.*

Proof Applying Lemmas 26 and 27 we get that with probability of at least $1-\delta$,

$$\hat{n} \leq O\left(\frac{d^2}{\lambda}\log\frac{dm}{\delta}\right).$$

Any balanced axis-aligned rectangle belongs to the class of all axis-aligned rectangles. Therefore, the coverage of CSS for the class of balanced axis-aligned rectangles is bounded bellow by the coverage of the class of axis-aligned rectangles. Applying Lemma 29, and assuming $m \geq d$, we obtain,

$$\gamma(\mathcal{H},\hat{n}) \quad \leq \quad O\left(d\frac{d^2}{\lambda}\log\frac{dm}{\delta}\log\left(\frac{d^2}{\lambda}\log\frac{dm}{\delta}\right)\right) \leq O\left(\frac{d^3}{\lambda}\log^2\frac{dm}{\lambda\delta}\right).$$

Applying Theorem 15 completes the proof. ■

6.2 Disagreement Coefficient

In this section we show interesting relations between the disagreement coefficient and coverage bounds in perfect selective classification. We begin by defining, for an hypothesis $h \in \mathcal{H}$, the set of all hypotheses that are r-close to h.

Definition 31 (Hanneke, 2011b, p.337) *For any hypothesis $h \in \mathcal{H}$, distribution P over $\mathcal{X}$, and $r > 0$, define the set $B(h,r)$ of all hypotheses that reside in a ball of radius r around h,*

$$B(h,r) \triangleq \left\{h' \in \mathcal{H} : \Pr_{X\sim P}\left\{h'(X) \neq h(X)\right\} \leq r\right\}.$$

Theorem 32 (Vapnik and Chervonenkis, 1971; Anthony and Bartlett, 1999, p.53) *Let $\mathcal{H}$ be a hypothesis class with VC-dimension d. For any probability distribution P on $\mathcal{X}\times\{\pm 1\}$, with probability of at least $1-\delta$ over the choice of S_m, any hypothesis $h \in \mathcal{H}$ consistent with S_m satisfies*

$$R(h) \leq \eta(d,m,\delta) \triangleq \frac{2}{m}\left[d\ln\frac{2em}{d} + \ln\frac{2}{\delta}\right].$$

For any $G \subseteq \mathcal{H}$ and distribution P we denote by ΔG the volume of the disagreement region of G,

$$\Delta G \triangleq \Pr\{DIS(G)\}.$$

Definition 33 (Disagreement coefficient, Hanneke, 2009) *Let $\varepsilon \geq 0$. The disagreement coefficient of the hypothesis class $\mathcal{H}$ with respect to the target distribution P is*

$$\theta(\varepsilon) \triangleq \theta_{h^*}(\varepsilon) = \sup_{r>\varepsilon} \frac{\Delta B(h^*, r)}{r}.$$

The following theorem formulates an intimate relation between active learning (disagreement coefficient) and selective classification.

Theorem 34 *Let $\mathcal{H}$ be an hypothesis class with VC-dimension d, P an unknown distribution, $\varepsilon \geq 0$, and $\theta(\varepsilon)$, the corresponding disagreement coefficient. Let (h,g) be a perfect selective classifier (CSS, see Section 2.3). Then, $R(h,g) = 0$, and for any $0 \leq \delta \leq 1$, with probability of at least $1-\delta$,*

$$\Phi(h,g) \geq 1 - \theta(\varepsilon) \cdot \max\{\eta(d,m,\delta), \varepsilon\}.$$

Proof Clearly, $R(h,g) = 0$, and it remains to prove the coverage bound. By Theorem 32, with probability of at least $1-\delta$,

$$\forall h \in VS_{\mathcal{H},S_m} \quad R(h) \leq \eta(d,m,\delta) \leq \max\{\eta(d,m,\delta), \varepsilon\}.$$

Therefore,

$$VS_{\mathcal{H},S_m} \subseteq B(h^*, \max\{\eta(d,m,\delta), \varepsilon\}),$$
$$\Delta VS_{\mathcal{H},S_m} \leq \Delta B(h^*, \max\{\eta(d,m,\delta), \varepsilon\}).$$

By Definition 33, for any $r' > \varepsilon$,

$$\Delta B(h^*, r') \leq \theta(\varepsilon) r'.$$

Thus, the proof is complete by recalling that

$$\Phi(h,g) = 1 - \Delta VS_{\mathcal{H},S_m}.$$

■

Theorem 34 tells us that whenever our learning problem (specified by the pair $(\mathcal{H},P)$) has a disagreement coefficient that grows slowly with respect to $1/\varepsilon$, it can be (perfectly) selectively learned with a "fast" coverage bound. Consequently, through Theorem 9 we also know that in each case where there exists a disagreement coefficient that grows slowly with respect to $1/\varepsilon$, active learning with a fast rate can also be deduced directly through a reduction from perfect selective classification. It follows that as far as fast rates in active learning are concerned, whatever can be accomplished by bounding the disagreement coefficient, can be accomplished also using perfect selective classification. This result is summarized in the following corollary.

Corollary 35 *Let $\mathcal{H}$ be an hypothesis class with VC-dimension d, P an unknown distribution, and $\theta(\varepsilon)$, the corresponding disagreement coefficient. If $\theta(\varepsilon) = O(polylog(1/\varepsilon))$, there exists a coverage bound such that an application of Theorem 7 ensures that $(\mathcal{H},P)$ is actively learnable with exponential label complexity speedup.*

Proof The proof is established by straightforward applications of Theorems 34 with $\varepsilon = 1/m$ and 9. ■

The following result, due to Hanneke (2011a), implies a coverage upper bound for CSS.

Lemma 36 (Hanneke, 2011a, Proof of Lemma 47) *Let $\mathcal{H}$ be an hypothesis class, P an unknown distribution, and $r \in (0,1)$. Then,*

$$\mathbf{E}_P \Delta D_m \geq (1-r)^m \Delta B(h^*, r),$$

where

$$D_m \triangleq VS_{\mathcal{H},S_m} \cap B(h^*, r). \tag{3}$$

Theorem 37 (Coverage upper bound) *Let $\mathcal{H}$ be an hypothesis class, P an unknown distribution, and $\delta \in (0,1)$. Then, for any $r \in (0,1)$, $1 > \alpha > \delta$,*

$$B_\Phi(\mathcal{H}, \delta, m) \leq 1 - \frac{(1-r)^m - \alpha}{1-\alpha} \Delta B(h^*, r),$$

where $B_\Phi(\mathcal{H}, \delta, m)$ is any coverage bound.

Proof Recalling the definition of D_m (3), clearly $D_m \subseteq VS_{\mathcal{H},S_m}$ and $D_m \subseteq B(h^*, r)$. These inclusions imply (respectively), by the definition of disagreement set,

$$\Delta D_m \leq \Delta VS_{\mathcal{H},S_m}, \text{ and } \Delta D_m \leq \Delta B(h^*, r). \tag{4}$$

Using Markov's inequality (in inequality (5) of the following derivation) and applying (4) (in equality (6)), we thus have,

$$\begin{aligned}
& Pr\left\{\Delta VS_{\mathcal{H},S_m} \leq \frac{(1-r)^m - \alpha}{1-\alpha} \Delta B(h^*, r)\right\} \leq Pr\left\{\Delta D_m \leq \frac{(1-r)^m - \alpha}{1-\alpha} \Delta B(h^*, r)\right\} \\
= \;& Pr\left\{\Delta B(h^*, r) - \Delta D_m \geq \frac{1-(1-r)^m}{1-\alpha} \Delta B(h^*, r)\right\} \\
\leq \;& Pr\left\{|\Delta B(h^*, r) - \Delta D_m| \geq \frac{1-(1-r)^m}{1-\alpha} \Delta B(h^*, r)\right\} \\
\leq \;& (1-\alpha) \cdot \frac{\mathbf{E}\{|\Delta B(h^*, r) - \Delta D_m|\}}{(1-(1-r)^m) \Delta B(h^*, r)} \qquad (5) \\
= \;& (1-\alpha) \cdot \frac{\Delta B(h^*, r) - \mathbf{E}\Delta D_m}{(1-(1-r)^m) \Delta B(h^*, r)}. \qquad (6)
\end{aligned}$$

Applying Lemma 36 we therefore obtain,

$$\leq \quad (1-\alpha) \cdot \frac{\Delta B(h^*, r) - (1-r)^m \Delta B(h^*, r)}{(1-(1-r)^m) \Delta B(h^*, r)} = 1 - \alpha < 1 - \delta.$$

Observing that for any coverage bound,

$$Pr\left\{\Delta VS_{\mathcal{H},S_m} \leq 1 - B_\Phi(\mathcal{H}, \delta, m)\right\} \geq 1 - \delta,$$

completes the proof. ■

Corollary 38 *Let $\mathcal{H}$ be an hypothesis class, P an unknown distribution, and $\delta \in (0,1/8)$. Then for any $m \geq 2$,*

$$B_\Phi(\mathcal{H},\delta,m) \leq 1 - \frac{1}{7}\Delta B\left(h^*, \frac{1}{m}\right),$$

where $B_\Phi(\mathcal{H},\delta,m)$ is any coverage bound.

Proof The proof is established by a straightforward application of Theorem 37 with $\alpha = 1/8$ and $r = 1/m$. ■

With Corollary 38 we can bound the disagreement coefficient for settings whose coverage bound is known.

Corollary 39 *Let $\mathcal{H}$ be an hypothesis class, P an unknown distribution, and $B_\Phi(\mathcal{H},\delta,m)$ a coverage bound. Then the disagreement coefficient is bounded by,*

$$\theta(\varepsilon) \leq \max\left\{\sup_{r\in(\varepsilon,1/2)} 7\cdot\frac{1-B_\Phi(\mathcal{H},1/9,\lfloor 1/r\rfloor)}{r}, 2\right\}.$$

Proof Applying Corollary 38 we get that for any $r \in (0,1/2)$,

$$\frac{\Delta B(h^*,r)}{r} \leq \frac{\Delta B(h^*,1/\lfloor 1/r\rfloor)}{r} \leq 7\cdot\frac{1-B_\Phi(\mathcal{H},1/9,\lfloor 1/r\rfloor)}{r}.$$

Therefore,

$$\theta(\varepsilon) = \sup_{r>\varepsilon}\frac{\Delta B(h^*,r)}{r} \leq \max\left\{\sup_{r\in(\varepsilon,1/2)} 7\cdot\frac{1-B_\Phi(\mathcal{H},1/9,\lfloor 1/r\rfloor)}{r}, 2\right\}.$$

■

Corollary 40 *Let $\mathcal{H}$ be the class of all linear binary classifiers in $\mathbb{R}^d$, and let the underlying distribution be any mixture of a fixed number of Gaussians in $\mathbb{R}^d$. Then*

$$\theta(\varepsilon) \leq O\left(polylog\left(\frac{1}{\varepsilon}\right)\right).$$

Proof Applying Corollary 39 together with inequality 2 we get that

$$\begin{aligned}\theta(\varepsilon) &\leq \max\left\{\sup_{r\in(\varepsilon,1/2)} 7\cdot\frac{1-B_\Phi(\mathcal{H},1/9,\lfloor 1/r\rfloor)}{r}, 2\right\} \\ &\leq \max\left\{\sup_{r\in(\varepsilon,1/2)} \frac{7}{r}\cdot O\left(\frac{(\log\lfloor 1/r\rfloor)^{d^2}}{\lfloor 1/r\rfloor}\cdot 9^{\frac{d+3}{2}}\right), 2\right\} \leq O\left(\left(log\frac{1}{\varepsilon}\right)^{d^2}\right).\end{aligned}$$

■

7. Concluding Remarks

For quite a few years, since its inception, the theory of target-independent bounds for noise-free active learning managed to handle relatively simple settings, mostly revolving around homogeneous linear classifiers under the uniform distribution over the sphere. It is likely that this distributional uniformity assumption was often adapted to simplify analyses. However, it was shown by Dasgupta (2005) that under this distribution, exponential speed up cannot be achieved when considering general (non homogeneous) linear classifiers.

The reason for this behavior is related to the two tasks that a good active learner should successfully accomplish: *exploration* and *exploitation*. Intuitively (and oversimplifying things) exploration is the task of obtaining at least one sample in each class, and exploitation is the process of refining the decision boundary by requesting labels of points around the boundary. Dasgupta showed that exploration cannot be achieved fast enough under the uniform distribution on the sphere. The source of this difficulty is the fact that under this distribution all training points reside on their convex hull. In general, the speed of exploration (using linear classifiers) depends on the size (number of vertices) of the convex hull of the training set. When using homogeneous linear classifiers, exploration is trivially achieved (under the uniform distribution) and exploitation can achieve exponential speedup.

So why in the *non-verifiable* model (Balcan et al., 2008) it is possible to achieve exponential speedup even when using non homogeneous linear classifiers under the uniform distribution? The answer is that in the non-verifiable model, label complexity attributed to exploration is encapsulated in a target-dependent "constant." Specifically, in Balcan et al. (2008) this constant is explicitly defined to be the probability mass of the minority class. Indeed, in certain noise free settings using linear classifiers, where the minority class is large enough, exploration is a non issue. In general, however, exploration is a major bottleneck in practical active learning (Baram et al., 2004; Begleiter et al., 2008). The present results show how exponential speedup can be achieved, including exploration, when using different (and perhaps more natural) distributions.

With these good news, a somewhat pessimistic picture arises from the lower bound we obtained for the exponential dependency on the dimension d. This negative result is not restricted to stream-based active learning and readily applies also to the pool-based model. While the bound is only asymptotic, we conjecture that it also holds for finite samples. Moreover, we believe that within the stream- or pool-based settings a similar statement should hold true for any active learning method (and not necessarily CAL-based querying strategies). This result indicates that when performing noise free active learning of linear classifiers, aggressive feature selection is beneficial for exploration speedup. We note, however, that it remains open whether a slowdown exponent of d (rather than d^2) is achievable.

We have exposed interesting relations of the present technique to well known complexity measures for active learning, namely, the teaching dimension and the disagreement coefficient. These developments were facilitated by observations made by Hanneke on the teaching dimension and the disagreement coefficient. These relations gave rise to further observations on active learning, which are discussed in Section 6 and include exponential speedup for balanced axis-aligned rectangles. Finally, we note that the intimate relation between selective classification and the disagreement coefficient was recently exposed in another result for selective classification where the disagreement coefficient emerged as a dominating factor in a coverage bound for agnostic selective classification (El-Yaniv and Wiener, 2011).

Acknowledgments

We thank the anonymous reviewers for their good comments. This paper particularly benefited from insightful observations made by one of the reviewers, which are summarized in Section 6, including the proof of Theorem 37 and the link between our $\hat{n}$ and the extended teaching dimension (Lemmas 26 and 27).

Appendix A.

Lemma 41 *For any $m \geq 3$, $a \geq 1$, $b \geq 1$ we get*

$$\sum_{i=1}^{m} \left(\frac{\ln^a (bi)}{i} \right) < \frac{4}{a} \ln^{a+1}(b(m+1)).$$

Proof Setting $f(x) \triangleq \frac{\ln^a (bx)}{x}$, we have

$$\frac{df}{dx} = (a - \ln bx) \cdot \frac{\ln^{a-1}(bx)}{x^2}.$$

Therefore, f is monotonically increasing when $x < e^a/b$, monotonically decreasing function when $x \geq e^a/b$ and its attains its maximum at $x = e^a/b$. Consequently, for $i < e^a/b - 1$, or $i \geq e^a/b + 1$,

$$f(i) \leq \int_{x=i-1}^{i+1} f(x)dx.$$

For $e^a/b - 1 \leq i < e^a/b + 1$,

$$f(i) \leq f(e^a/b) = b \left(\frac{a}{e} \right)^a \leq a^a. \tag{7}$$

Therefore, if $m < e^a - 1$ we have,

$$\sum_{i=1}^{m} f(i) = \ln^a (b) + \sum_{i=2}^{m} f(i) < 2 \cdot \int_{x=1}^{m+1} f(x)dx \leq \frac{2}{a+1} \ln^{a+1}(b(m+1)).$$

Otherwise, $m \geq e^a/b$, in which case we overcome the change of slope by adding twice the (upper bound on the) maximal value (7),

$$\begin{aligned} \sum_{i=1}^{m} f(i) &< \frac{2}{a+1} \ln^{a+1}(b(m+1)) + 2a^a = \frac{2}{a+1} \ln^{a+1}(b(m+1)) + \frac{2}{a} a^{a+1} \\ &\leq \frac{2}{a+1} \ln^{a+1}(b(m+1)) + \frac{2}{a} \ln^{a+1} bm \leq \frac{4}{a} \ln^{a+1}(b(m+1)). \end{aligned}$$

■

Appendix B. Alternative Proof of Lemma 6 Using Super Martingales

Define $W_k \triangleq \sum_{i=1}^{k}(Z_i - b_i)$. We assume that with probability of at least $1-\delta/2$, $\Pr\{Z_i|Z_1,\ldots,Z_{i-1}\} \leq b_i$, simultaneously for all i. Since Z_i is a binary random variable it is easy to see that (w.h.p.),

$$E_{Z_i}\{W_i|Z_1,\ldots,Z_{i-1}\} = \Pr\{Z_i|Z_1,\ldots,Z_{i-1}\} - b_i + W_{i-1} \leq W_{i-1},$$

and the sequence $W_1^m \triangleq W_1,\ldots,W_m$ is a super-martingale with high probability. We apply the following theorem by McDiarmid that refers to martingales (but can be shown to apply to super-martingales, by following its original proof).

Theorem 42 (McDiarmid, 1998, Theorem 3.12) *Let $Y_1,\ldots,Y_n$ be a martingale difference sequence with $-a_k \leq Y_k \leq 1-a_k$ for each k; let $A = \frac{1}{n}\sum a_k$. Then, for any $\varepsilon > 0$,*

$$\Pr\left\{\sum Y_k \geq An\varepsilon\right\} \leq \exp\left(-[(1+\varepsilon)\ln(1+\varepsilon)-\varepsilon]An\right) \leq \exp\left(-\frac{An\varepsilon^2}{2(1+\varepsilon/3)}\right).$$

In our case, $Y_k = W_k - W_{k-1} = Z_k - b_k \leq 1-b_k$ and we apply the (revised) theorem with $a_k \triangleq b_k$ and $An \triangleq \sum b_k \triangleq B$. We thus obtain, for any $0 < \varepsilon < 1$,

$$\Pr\left\{\sum Z_k \geq B + B\varepsilon\right\} \leq \exp\left(-\frac{B\varepsilon^2}{2(1+\varepsilon/3)}\right).$$

Equating the right-hand side to $\delta/2$, we obtain

$$\begin{aligned}
\varepsilon &= \left(\frac{2}{3}\ln\frac{2}{\delta} \pm \sqrt{\frac{4}{9}\ln^2\frac{2}{\delta} + 8B\ln\frac{2}{\delta}}\right)/2B \\
&\leq \left(\frac{1}{3}\ln\frac{2}{\delta} + \sqrt{\frac{1}{9}\ln^2\frac{2}{\delta}} + \sqrt{2B\ln\frac{2}{\delta}}\right)/B \\
&= \left(\frac{2}{3}\ln\frac{2}{\delta} + \sqrt{2B\ln\frac{2}{\delta}}\right)/B.
\end{aligned}$$

Applying the union bound completes the proof.

References

M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations*. Cambridge University Press, 1999.

L. Atlas, D. Cohn, R. Ladner, A.M. El-Sharkawi, and R.J. Marks. Training connectionist networks with queries and selective sampling. In *Neural Information Processing Systems (NIPS)*, pages 566–573, 1990.

M.F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *21st Annual Conference on Learning Theory (COLT)*, pages 45–56, 2008.

Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, 2004.

P.L. Bartlett and M.H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

R. Begleiter, R. El-Yaniv, and D. Pechyony. Repairing self-confident active-transductive learners using systematic exploration. *Pattern Recognition Letters*, 29(9):1245–1251, 2008.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36, 1989.

C.K. Chow. An optimum character recognition system using decision function. *IEEE Transactions on Computers*, 6(4):247–254, 1957.

C.K. Chow. On optimum recognition error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–36, 1970.

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

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, pages 235–242, 2005.

S. Dasgupta, A. Tauman Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *Neural Information Processing Systems (NIPS)*, 2011.

S. Fine, R. Gilad-Bachrach, and E. Shamir. Query by committee, linear separation and random walks. *Theoretical Computer Science*, 284(1):25–51, 2002.

Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Information, prediction, and Query by Committee. In *Advances in Neural Information Processing Systems (NIPS) 5*, pages 483–490, 1993.

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

Y. Freund, Y. Mansour, and R.E. Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, 32(4):1698–1722, 2004.

E. Friedman. Active learning for smooth problems. *In Proceedings of the* 22^{nd} *Annual Conference on Learning Theory (COLT)*, 2009.

R. Gilad-Bachrach. *To PAC and Beyond*. PhD thesis, the Hebrew University of Jerusalem, 2007.

S. Goldman and M. Kearns. On the complexity of teaching. *JCSS: Journal of Computer and System Sciences*, 50, 1995.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007a.

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, volume 4539 of *Lecture Notes in Artificial Intelligence*, pages 66–81, 2007b.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *CoRR*, abs/1108.1766, 2011a. URL `http://arxiv.org/abs/1108.1766`. informal publication.

S. Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 37(1):333–361, 2011b.

T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1995.

R. Herbei and M.H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34(4):709–721, 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

D. Hug and M. Reitzner. Gaussian polytopes: variances and limit theorems, June 2005.

D. Hug, G. O. Munsonious, and M. Reitzner. Asymptotic mean values of Gaussian polytopes. *Beiträge Algebra Geom.*, 45:531–548, 2004.

C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16, pages 195–248. Springer-Verlag, 1998.

T. Mitchell. Version spaces: a candidate elimination approach to rule learning. In *IJCAI'77: Proceedings of the 5th international joint conference on Artificial Intelligence*, pages 305–310, 1977.

H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning theory (COLT)*, pages 287–294, 1992.

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

M.H. Wegkamp. Lasso type classifiers with a reject option. *Electronic Journal of Statistics*, 1: 155–168, 2007.