Journal of Machine Learning Research 16 (2015) 713-745 Submitted 4/14; Published 4/15

A Compression Technique for Analyzing Disagreement-Based Active Learning

Yair Wiener YAIR.WIENER@GMAIL.COM
Computer Science Department
Technion – Israel Institute of Technology
Haifa 32000, Israel

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Princeton, NJ 08542 USA

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

Editor: Sanjoy Dasgupta

Abstract

We introduce a new and improved characterization of the label complexity of disagreement-based active learning, in which the leading quantity is the *version space compression set size*. This quantity is defined as the size of the smallest subset of the training data that induces the same version space. We show various applications of the new characterization, including a tight analysis of CAL and refined label complexity bounds for linear separators under mixtures of Gaussians and axis-aligned rectangles under product densities. The version space compression set size, as well as the new characterization of the label complexity, can be naturally extended to agnostic learning problems, for which we show new speedup results for two well known active learning algorithms.

Keywords: active learning, selective sampling, sequential design, statistical learning theory, PAC learning, sample complexity, selective prediction

1. Introduction

Active learning is a learning paradigm allowing the learner to sequentially request the target labels of selected instances from a pool or stream of unlabeled data.[1] The key question in the theoretical analysis of active learning is how many label requests are sufficient to learn the labeling function to a specified accuracy, a quantity known as the *label complexity*. Among the many recent advances in the theory of active learning, perhaps the most well-studied technique has been the *disagreement-based* approach, initiated by Cohn, Atlas, and Ladner (1994), and further advanced in numerous articles (e.g., Balcan, Beygelzimer, and Langford, 2009; Dasgupta, Hsu, and Monteleoni, 2007; Beygelzimer, Dasgupta, and Langford, 2009; Beygelzimer, Hsu, Langford, and Zhang, 2010; Koltchinskii, 2010; Hanneke, 2012; Hanneke and Yang, 2012). The basic strategy in disagreement-based active learning is to sequentially process the unlabeled examples, and for each example, the algorithm requests its label if and only if the value of the optimal classifier's classification on that point cannot be inferred from information already obtained.

1. Any active learning technique for streaming data can be used in pool-based models but not vice versa

One attractive feature of this approach is that its simplicity makes it amenable to thorough theoretical analysis, and numerous theoretical guarantees on the performance of variants of this strategy under various conditions have appeared in the literature (see e.g., Balcan, Beygelzimer, and Langford, 2009; Hanneke, 2007a; Dasgupta, Hsu, and Monteleoni, 2007; Balcan, Broder, and Zhang, 2007; Beygelzimer, Dasgupta, and Langford, 2009; Friedman, 2009; Balcan, Hanneke, and Vaughan, 2010; Hanneke, 2011; Koltchinskii, 2010; Beygelzimer, Hsu, Langford, and Zhang, 2010; Hsu, 2010; Hanneke, 2012; El-Yaniv and Wiener, 2012; Hanneke and Yang, 2012; Hanneke, 2014). The majority of these results formulate bounds on the label complexity in terms of a complexity measure known as the *disagreement coefficient* (Hanneke, 2007a), which we define below. A notable exception to this is the recent work of El-Yaniv and Wiener (2012), rooted in the related topic of selective prediction (El-Yaniv and Wiener, 2010; Wiener and El-Yaniv, 2012; Wiener, 2013; Wiener and El-Yaniv, 2015), which instead bounds the label complexity in terms of two complexity measures called the *characterizing set complexity* and the *version space compression set size* (El-Yaniv and Wiener, 2010). In the current literature, the above are the only known general techniques for the analysis of disagreement-based active learning.

In the present article, we present a new characterization of the label complexity of disagreement-based active learning. The leading quantity in our characterization is the *version space compression set size* of El-Yaniv and Wiener (2012, 2010); Wiener (2013), which corresponds to the size of the smallest subset of the training set that induces the same version space as the entire training set. This complexity measure was shown by El-Yaniv and Wiener (2012) to be a special case of the extended teaching dimension of Hanneke (2007b).

The new characterization improves upon the two prior techniques in some cases. For a noiseless setting (the realizable case), we show that the label complexity results derived from this new technique are *tight* up to logarithmic factors. This was not true of either of the previous techniques; as we discuss in Appendix B, the known upper bounds in the literature expressed in terms of these other complexity measures are sometimes off by a factor of the VC dimension. Moreover, the new method significantly simplifies the recent technique of Wiener (2013); El-Yaniv and Wiener (2012, 2010) by completely eliminating the need for the characterizing set complexity measure.

Interestingly, interpreted as an upper bound on the label complexity of active learning in general, the upper bounds presented here also reflect improvements over a bound of Hanneke (2007b), which is also expressed in terms of (a target-independent variant of) this same complexity measure: specifically, reducing the bound by roughly a factor of the VC dimension compared to that result. In addition to these results on the label complexity, we also relate the version space compression set size to the disagreement coefficient, essentially showing that they are always within a factor of the VC dimension of each other (with additional logarithmic factors).

We apply this new technique to derive new results for two learning problems: namely, linear separators under mixtures of Gaussians, and axis-aligned hyperrectangles under product densities. We derive bounds on the version space compression set size for each of these. Thus, using our results relating the version space compression set size to the label complexity, we arrive at bounds on the label complexity of disagreement-based active learning for these problems, which represent significant refinements of the best results in the prior literature on these settings.

While the version space compression set size is initially defined for noiseless (realizable) learning problems that have a version space, it can be naturally extended to an agnostic setting, and the new technique applies to noisy, agnostic problems as well. This surprising result, which was motivated by related observations of Hanneke (2014); Wiener (2013), is allowed through bounds on the disagreement coefficient in terms of the version space compression set size, and the applicability of the disagreement coefficient to both the realizable and agnostic settings. We formulate this generalization in Section 6 and present new sample complexity results for known active learning algorithms, including the disagreement-based methods of Dasgupta, Hsu, and Monteleoni (2007) and Hanneke (2012). These results tighten the bounds of Wiener (2013) using the new technique.

2. Preliminary Definitions

Let $\mathcal{X}$ denote a set, called the *instance space*, and let $\mathcal{Y} \triangleq \{-1,+1\}$, called the *label space*. A *classifier* is a measurable function $h : \mathcal{X} \to \mathcal{Y}$. Throughout, we fix a set $\mathcal{F}$ of classifiers, called the *concept space*, and denote by d the VC dimension of $\mathcal{F}$ (Vapnik and Chervonenkis, 1971; Vapnik, 1998). We also fix an arbitrary probability measure P over $\mathcal{X} \times \mathcal{Y}$, called the *data distribution*. Aside from Section 6, we make the assumption that $\exists f^* \in \mathcal{F}$ with $\mathbb{P}(Y = f^*(x)|X = x) = 1$ for all $x \in \mathcal{X}$, where $(X,Y) \sim P$; this is known as the *realizable case*, and f^* is known as the *target function*. For any classifier h, define its *error rate* $\mathrm{er}(h) \triangleq P((x,y) : h(x) \neq y)$; note that $\mathrm{er}(f^*) = 0$.

For any set $\mathcal{H}$ of classifiers, define the *region of disagreement*

$$\mathrm{DIS}(\mathcal{H}) \triangleq \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}.$$

Also define $\Delta\mathcal{H} \triangleq P(\mathrm{DIS}(\mathcal{H}) \times \mathcal{Y})$, the marginal probability of the region of disagreement.

Let $S_\infty \triangleq \{(x_1,y_1),(x_2,y_2),\ldots\}$ be a sequence of i.i.d. P-distributed random variables, and for each $m \in \mathbb{N}$, denote by $S_m \triangleq \{(x_1,y_1),\ldots,(x_m,y_m)\}$.[2] For any $m \in \mathbb{N} \cup \{0\}$, and any $S \in (\mathcal{X} \times \mathcal{Y})^m$, define the *version space* $\mathrm{VS}_{\mathcal{F},S} \triangleq \{h \in \mathcal{F} : \forall (x,y) \in S, h(x) = y\}$ (Mitchell, 1977). The following definition will be central in our results below.

Definition 1 (Version Space Compression Set Size) *For any $m \in \mathbb{N} \cup \{0\}$ and any $S \in (\mathcal{X} \times \mathcal{Y})^m$, the* version space compression set *$\hat{C}_S$ is a smallest subset of S satisfying $\mathrm{VS}_{\mathcal{F},\hat{C}_S} = \mathrm{VS}_{\mathcal{F},S}$. The* version space compression set size *is defined to be $\hat{n}(\mathcal{F},S) \triangleq |\hat{C}_S|$. In the special cases where $\mathcal{F}$ and perhaps $S = S_m$ are obvious from the context, we abbreviate $\hat{n} \triangleq \hat{n}(S_m) \triangleq \hat{n}(\mathcal{F},S_m)$.*

Note that the value $\hat{n}(\mathcal{F},S)$ is unique for any S, and $\hat{n}(S_m)$ is, obviously, a random number that depends on the (random) sample S_m. The quantity $\hat{n}(S_m)$ has been studied under at least two names in the prior literature. Drawing motivation from the work on Exact learning with Membership Queries (Hegedüs, 1995; Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins, 1996), which extends ideas from Goldman and Kearns (1995) on the complexity of teaching, the quantity $\hat{n}(S_m)$ was introduced in the work of Hanneke (2007b) as the *extended teaching dimension* of the classifier f^* on the space $\{x_1,\ldots,x_m\}$ with respect to the set $\mathcal{F}[\{x_1,\ldots,x_m\}] \triangleq \{x_i \mapsto h(x_i) : h \in \mathcal{F}\}$ of distinct classifications of $\{x_1,\ldots,x_m\}$ realized by $\mathcal{F}$; in this context, the set $\hat{C}_{S_m}$ is known as a *minimal specifying set* of f^* on $\{x_1,\ldots,x_m\}$ with respect to $\mathcal{F}[\{x_1,\ldots,x_m\}]$. The quantity $\hat{n}(S_m)$ was independently discovered by El-Yaniv and Wiener (2010) in the context of selective classification, which is the source of the compression set terminology introduced above; we adopt this terminology throughout the present article. See the work of El-Yaniv and Wiener (2012) for a formal proof of the equivalence of these two notions.

It will also be useful to define minimal confidence bounds on certain quantities, as follows.

2. Note that, in the realizable case, $y_i = f^*(x_i)$ for all i with probability 1. For simplicity, we will suppose these equalities hold throughout our discussion of the realizable case.

Definition 2 (Version Space Compression Set Size Minimal Bound) *For any* $m \in \mathbb{N} \cup \{0\}$ *and* $\delta \in (0,1]$*, define the* version space compression set size minimal bound

$$\mathcal{B}_{\hat{n}}(m,\delta) \triangleq \min\left\{b \in \mathbb{N} \cup \{0\} : \mathbb{P}(\hat{n}(S_m) \le b) \ge 1-\delta\right\}.$$

Similarly, define the *version space disagreement region minimal bound*

$$\mathcal{B}_{\Delta}(m,\delta) \triangleq \min\left\{t \in [0,1] : \mathbb{P}(\Delta \mathrm{VS}_{\mathcal{F},S_m} \le t) \ge 1-\delta\right\}.$$

In both cases, the quantities implicitly also depend on $\mathcal{F}$ and P (which remain fixed throughout our analysis below), and the only random variables involved in these probabilities are the data S_m.

Most of the existing general results on disagreement-based active learning are expressed in terms of a quantity known as the *disagreement coefficient* (Hanneke, 2007a, 2009), defined as follows.

Definition 3 (Disagreement Coefficient) *For any classifier f and $r > 0$, define the r-ball centered at f as*

$$\mathrm{B}(f,r) \triangleq \{h \in \mathcal{F} : \Delta\{h,f\} \le r\},$$

and for any $r_0 \ge 0$, define the disagreement coefficient *of $\mathcal{F}$ with respect to P as*[3]

$$\theta(r_0) \triangleq \sup_{r>r_0} \frac{\Delta \mathrm{B}(f^*,r)}{r} \vee 1.$$

The disagreement coefficient was originally introduced to the active learning literature by Hanneke (2007a), and has been studied and bounded by a number of authors (see e.g., Hanneke, 2007a; Friedman, 2009; Wang, 2011; Hanneke, 2014; Balcan and Long, 2013). Similar quantities have also been studied in the passive learning literature, rooted in the work of Alexander (see e.g., Alexander, 1987; Giné and Koltchinskii, 2006).

Numerous recent results, many of which are surveyed by Hanneke (2014), exhibit bounds on the label complexity of disagreement-based active learning in terms of the disagreement coefficient. It is therefore of major interest to develop such bounds for specific cases of interest (i.e., for specific classes $\mathcal{F}$ and distributions P). In particular, any result showing $\theta(r_0) = o(1/r_0)$ indicates that disagreement-based active learning should asymptotically provide some advantage over passive learning for that $\mathcal{F}$ and P (Hanneke, 2012). We are particularly interested in scenarios in which $\theta(r_0) = O(\mathrm{polylog}(1/r_0))$, or even $\theta(r_0) = O(1)$, since these imply strong improvements over passive learning (Hanneke, 2007a, 2011).

There are several general results on the asymptotic behavior of the disagreement coefficient as $r_0 \to 0$, for interesting cases. For the class of linear separators in $\mathbb{R}^k$, perhaps the most general result to date is that the existence of a density function for the marginal distribution of P over $\mathcal{X}$ is sufficient to guarantee $\theta(r_0) = o(1/r_0)$ (Hanneke, 2014). That work also shows that, if the density is bounded and has bounded support, and the target separator passes through the support at a continuity point of the density, then $\theta(r_0) = O(1)$. In both of these cases, for $k \ge 2$, the specific dependence on r_0 in the little-o and the constant factors in the big-O will vary depending on the particular distribution P, and in particular, will depend on f^* (i.e., such bounds are *target-dependent*).

There are also several explicit, *target-independent* bounds on the disagreement coefficient in the literature. Perhaps the most well-known of these is for homogeneous linear separators in $\mathbb{R}^k$, where

3. We use the notation $a \vee b = \max\{a,b\}$.

the marginal distribution of P over $\mathcal{X}$ is confined to be the uniform distribution over the unit sphere, in which case $\theta(r_0)$ is known to be within a factor of 4 of $\min\{\pi\sqrt{k}, 1/r_0\}$ (Hanneke, 2007a). In the present paper, we are primarily focused on explicit, target-independent speedup bounds, though our abstract results can be used to derive bounds of either type.

3. Relating $\hat{n}$ and the Disagreement Coefficient

In this section, we show how to bound the disagreement coefficient in terms of $\mathcal{B}_{\hat{n}}(m,\delta)$. We also show the other direction and bound $\mathcal{B}_{\hat{n}}(m,\delta)$ in terms of the disagreement coefficient.

Theorem 4 *For any* $r_0 \in (0,1)$,

$$\theta(r_0) \leq \max\left\{ \max_{r \in (r_0,1)} 16\mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20}\right), 512 \right\}.$$

Proof We will prove that, for any $r \in (0,1)$,

$$\frac{\Delta \mathrm{B}(f^*,r)}{r} \leq \max\left\{ 16\mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20}\right), 512 \right\}. \tag{1}$$

The result then follows by taking the supremum of both sides over $r \in (r_0, 1)$.

Fix $r \in (0,1)$, let $m = \lceil 1/r \rceil$, and for $i \in \{1,\ldots,m\}$, define $S_{m\setminus i} = S_m \setminus \{(x_i,y_i)\}$. Also define $D_{m\setminus i} = \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}} \cap \mathrm{B}(f^*,r))$ and $\Delta_{m\setminus i} = \mathbb{P}(x_i \in D_{m\setminus i} | S_{m\setminus i}) = P(D_{m\setminus i} \times \mathcal{Y})$. If $\Delta\mathrm{B}(f^*,r)m \leq 512$, (1) clearly holds. Otherwise, suppose $\Delta\mathrm{B}(f^*,r)m > 512$. If $x_i \in \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})$, then we must have $(x_i,y_i) \in \hat{\mathcal{C}}_{S_m}$. So

$$\hat{n}(S_m) \geq \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})}(x_i).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\{\hat{n}(S_m) \leq (1/16)\Delta\mathrm{B}(f^*,r)m\} \\
&\leq \mathbb{P}\left\{ \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{m\setminus i}})}(x_i) \leq (1/16)\Delta\mathrm{B}(f^*,r)m \right\} \\
&\leq \mathbb{P}\left\{ \sum_{i=1}^{m} \mathbb{1}_{D_{m\setminus i}}(x_i) \leq (1/16)\Delta\mathrm{B}(f^*,r)m \right\} \\
&= \mathbb{P}\left\{ \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \mathbb{1}_{D_{m\setminus i}}(x_i) \geq \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - (1/16)\Delta\mathrm{B}(f^*,r)m \right\}.
\end{aligned}$$

Breaking the above event into two cases based on the value of $\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i)$, this last line equals

$$\begin{aligned}
&\mathbb{P}\Bigg\{\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \mathbb{1}_{D_{m\setminus i}}(x_i) \geq \\
&\qquad \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \frac{1}{16}\Delta\mathrm{B}(f^*,r)m, \quad \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < \frac{7}{8}\Delta\mathrm{B}(f^*,r)m\Bigg\} \\
&+\mathbb{P}\Bigg\{\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \mathbb{1}_{D_{m\setminus i}}(x_i) \geq \\
&\qquad \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \frac{1}{16}\Delta\mathrm{B}(f^*,r)m, \quad \sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) \geq \frac{7}{8}\Delta\mathrm{B}(f^*,r)m\Bigg\} \\
&\leq \mathbb{P}\Bigg\{\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < (7/8)\Delta\mathrm{B}(f^*,r)m\Bigg\} \\
&\quad +\mathbb{P}\Bigg\{\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \mathbb{1}_{D_{m\setminus i}}(x_i) \geq (13/16)\Delta\mathrm{B}(f^*,r)m\Bigg\}.
\end{aligned}$$

Since we are considering the case $\Delta\mathrm{B}(f^*,r)m > 512$, a Chernoff bound implies

$$\mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) < (7/8)\Delta\mathrm{B}(f^*,r)m\right) \leq \exp\{-\Delta\mathrm{B}(f^*,r)m/128\} < e^{-4}.$$

Furthermore, Markov's inequality implies

$$\mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r))}(x_i) - \mathbb{1}_{D_{m\setminus i}}(x_i) \geq (13/16)\Delta\mathrm{B}(f^*,r)m\right) \leq \frac{m\Delta\mathrm{B}(f^*,r) - \mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{D_{m\setminus i}}(x_i)\right]}{(13/16)m\Delta\mathrm{B}(f^*,r)}.$$

Since the x_i values are exchangeable,

$$\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{D_{m\setminus i}}(x_i)\right] = \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{D_{m\setminus i}}(x_i)\middle|S_{m\setminus i}\right]\right] = \sum_{i=1}^{m} \mathbb{E}\left[\Delta_{m\setminus i}\right] = m\mathbb{E}\left[\Delta_{m\setminus m}\right].$$

Hanneke (2012) proves that this is at least

$$m(1-r)^{m-1}\Delta\mathrm{B}(f^*,r).$$

In particular, when $\Delta\mathrm{B}(f^*,r)m > 512$, we must have $r < 1/511 < 1/2$, which implies $(1-r)^{\lceil 1/r\rceil -1} \geq 1/4$, so that we have

$$\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{D_{m\setminus i}}(x_i)\right] \geq (1/4)m\Delta\mathrm{B}(f^*,r).$$

Altogether, we have established that

$$\mathbb{P}\left(\hat{n}(S_m) \leq (1/16)\Delta\mathrm{B}(f^*,r)m\right) < \frac{m\Delta\mathrm{B}(f^*,r) - (1/4)m\Delta\mathrm{B}(f^*,r)}{(13/16)m\Delta\mathrm{B}(f^*,r)} + e^{-4} = \frac{12}{13} + e^{-4} < \frac{19}{20}.$$

Thus, since $\hat{n}(S_m) \leq \mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right)$ with probability at least $\frac{19}{20}$, we must have that

$$\mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) > (1/16)\Delta\mathrm{B}(f^*, r)m \geq (1/16)\frac{\Delta\mathrm{B}(f^*, r)}{r}.$$

■

The following Theorem, whose proof is given in Section 4, is a "converse" of Theorem 4, showing a bound on $\mathcal{B}_{\hat{n}}(m, \delta)$ in terms of the disagreement coefficient.

Theorem 5 *There is a finite universal constant $c > 0$ such that, $\forall r_0, \delta \in (0,1)$,*

$$\max_{r \in (r_0, 1)} \mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \delta\right) \leq c\theta(dr_0)\left(d \ln(e\theta(dr_0)) + \ln\left(\frac{\log_2(2/r_0)}{\delta}\right)\right)\log_2\left(\frac{2}{r_0}\right).$$

4. A Tight Analysis of CAL

The following algorithm is due to Cohn, Atlas, and Ladner (1994).

Algorithm: CAL(n)
0. $m \leftarrow 0$, $t \leftarrow 0$, $V_0 \leftarrow \mathcal{F}$
1. While $t < n$
2. $\quad m \leftarrow m + 1$
3. $\quad$ If $x_m \in \mathrm{DIS}(V_{m-1})$
4. $\quad\quad$ Request label y_m; let $V_m \leftarrow \{h \in V_{m-1} : h(x_m) = y_m\}$, $t \leftarrow t + 1$
5. $\quad$ Else $V_m \leftarrow V_{m-1}$
6. Return any $\hat{h} \in V_m$

One particularly attractive feature of this algorithm is that it maintains the invariant that $V_m = \mathrm{VS}_{\mathcal{F}, S_m}$ for all values of m it obtains (since, if $V_{m-1} = \mathrm{VS}_{\mathcal{F}, S_{m-1}}$, then $f^* \in V_{m-1}$, so any point $x_m \notin \mathrm{DIS}(V_{m-1})$ has $\{h \in V_{m-1} : h(x_m) = y_m\} = \{h \in V_{m-1} : h(x_m) = f^*(x_m)\} = V_{m-1}$ anyway). To analyze this method, we first define, for every $m \in \mathbb{N}$,

$$N(m; S_m) = \sum_{t=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F}, S_{t-1}})}(x_t),$$

which counts the number of labels requested by CAL among the first m data points (assuming it does not halt first). The following result provides data-dependent upper and lower bounds on this important quantity, which will be useful in establishing label complexity bounds for CAL below.

Lemma 6

$$\max_{t \leq m} \hat{n}(S_t) \leq N(m; S_m),$$

and with probability at least $1 - \delta$,

$$N(m; S_m) \leq \max_{t \in \{2^i : i \in \{0, \ldots, \lfloor \log_2(m) \rfloor\}\}} \left(55\hat{n}(S_t)\ln\left(\frac{et}{\hat{n}(S_t)}\right) + 24\ln\left(\frac{4\log_2(2m)}{\delta}\right)\right)\log_2(2m).$$

Since the upper and lower bounds on $N(m;S_m)$ in Lemma 6 require access to the *labels* of the data, they are not as much interesting for practice as they are for their theoretical significance. In particular, they will allow us to derive new distribution-dependent bounds on the performance of CAL below (Theorems 9 and 10). Lemma 6 is also of some *conceptual* significance, as it shows a direct and fairly-tight connection between the behavior of CAL and the size of the version space compression set.

The proof of the upper bound on $N(m;S_m)$ relies on the following two lemmas. The first lemma (Lemma 7) is implied by a classical compression bound of Littlestone and Warmuth (1986), and provides a high-confidence bound on the probability measure of a set, given that it has zero empirical frequency and is specified by a small number of samples. For completeness, we include a proof of this result below: a variant of the original argument of Littlestone and Warmuth (1986).[4]

Lemma 7 (Compression; Littlestone and Warmuth, 1986) *For any $\delta \in (0,1)$, any collection $\mathbb{D}$ of measurable sets $D \subseteq \mathcal{X} \times \mathcal{Y}$, any $m \in \mathbb{N}$ and $n \in \mathbb{N} \cup \{0\}$ with $n \leq m$, and any permutation-invariant function $\phi_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{D}$, with probability of at least $1-\delta$ over draw of S_m, every distinct $i_1, \ldots, i_n \in \{1, \ldots, m\}$ with $S_m \cap \phi_n((x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n})) = \emptyset$ satisfies*[5]

$$P(\phi_n((x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n}))) \leq \frac{1}{m-n}\left(n \ln\left(\frac{em}{n}\right) + \ln\left(\frac{1}{\delta}\right)\right). \tag{2}$$

Proof Let $\varepsilon > 0$ denote the value of the right hand side of (2). The result trivially holds if $\varepsilon > 1$. For the remainder, consider the case $\varepsilon \leq 1$. Let I_n be the set of all sets of n distinct indices $\{i_1, \ldots, i_n\}$ from $\{1, \ldots, m\}$. Note that $|I_n| = \binom{m}{n}$. Given a labeled sample S_m and $\mathbf{i} = \{i_1, \ldots, i_n\} \in I_n$, denote by $S_m^{\mathbf{i}} = \{(x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n})\}$, and by $S_m^{-\mathbf{i}} = \{(x_i, y_i) : i \in \{1, \ldots, m\} \setminus \mathbf{i}\}$. Since ϕ_n is permutation-invariant, for any distinct $i_1, \ldots, i_n \in \{1, \ldots, m\}$, letting $\mathbf{i} = \{i_1, \ldots, i_n\}$ denote the unordered set of indices, we may denote $\phi_n(S_m^{\mathbf{i}}) = \phi_n((x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n}))$ without ambiguity. In particular, we have $\{\phi_n((x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n})) : i_1, \ldots, i_n \in \{1, \ldots, m\} \text{ distinct}\} = \{\phi_n(S_m^{\mathbf{i}}) : \mathbf{i} \in I_n\}$, so that it suffices to show that, with probability at least $1-\delta$, every $\mathbf{i} \in I_n$ with $S_m \cap \phi_n(S_m^{\mathbf{i}}) = \emptyset$ has $P(\phi_n(S_m^{\mathbf{i}})) \leq \varepsilon$.

Define the events $\omega(\mathbf{i}, m) = \left\{S_m \cap \phi_n(S_m^{\mathbf{i}}) = \emptyset\right\}$ and $\omega'(\mathbf{i}, m-n) = \left\{S_m^{-\mathbf{i}} \cap \phi_n(S_m^{\mathbf{i}}) = \emptyset\right\}$. Note that $\omega(\mathbf{i}, m) \subseteq \omega'(\mathbf{i}, m-n)$. Therefore, for each $\mathbf{i} \in I_n$, we have

$$\mathbb{P}\left(\left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\} \cap \omega(\mathbf{i}, m)\right) \leq \mathbb{P}\left(\left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\} \cap \omega'(\mathbf{i}, m-n)\right).$$

By the law of total probability and $\sigma(S_m^{\mathbf{i}})$-measurability of the event $\left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\}$, this equals

$$\mathbb{E}\left[\mathbb{P}\left(\left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\} \cap \omega'(\mathbf{i}, m-n) \middle| S_m^{\mathbf{i}}\right)\right] = \mathbb{E}\left[\mathbb{1}[P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon]\,\mathbb{P}\left(\omega'(\mathbf{i}, m-n) \middle| S_m^{\mathbf{i}}\right)\right].$$

Noting that $|S_m^{-\mathbf{i}} \cap \phi_n(S_m^{\mathbf{i}})|$ is conditionally Binomial$(m-n, P(\phi_n(S_m^{\mathbf{i}})))$ given $S_m^{\mathbf{i}}$, this equals

$$\mathbb{E}\left[\mathbb{1}[P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon]\left(1 - P(\phi_n(S_m^{\mathbf{i}}))\right)^{m-n}\right] \leq (1-\varepsilon)^{m-n} \leq e^{-\varepsilon(m-n)},$$

4. See also Section 5.2.1 of Herbrich (2002) for a very clear and concise proof of a similar result (beginning with the line above (5.15) there, for our purposes).
5. We define $0 \ln(1/0) = 0 \ln(\infty) = 0$.

where the last inequality is due to $1-\varepsilon \leq e^{-\varepsilon}$ (see e.g., Theorem A.101 of Herbrich, 2002). In the case $n=0$, this last expression equals δ, which establishes the result since $|I_0|=1$. Otherwise, if $n>0$, combining the above with a union bound, we have that

$$\mathbb{P}\left(\exists \mathbf{i} \in I_n : P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon \wedge S_m \cap \phi_n(S_m^{\mathbf{i}}) = \emptyset\right) = \mathbb{P}\left(\bigcup_{\mathbf{i}\in I_n} \left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\} \cap \omega(\mathbf{i},m)\right)$$
$$\leq \sum_{\mathbf{i}\in I_n} \mathbb{P}\left(\left\{P(\phi_n(S_m^{\mathbf{i}})) > \varepsilon\right\} \cap \omega(\mathbf{i},m)\right) \leq \sum_{\mathbf{i}\in I_n} e^{-\varepsilon(m-n)} = \binom{m}{n} e^{-\varepsilon(m-n)}.$$

Since $\binom{m}{n} \leq \left(\frac{em}{n}\right)^n$ (see e.g., Theorem A.105 of Herbrich, 2002), this last expression is at most $\left(\frac{em}{n}\right)^n e^{-\varepsilon(m-n)} = \delta$, which completes the proof. ■

The following, Lemma 8, will be used for proving Lemma 6 above. The lemma relies on Lemma 7 and provides a high-confidence bound on the probability of requesting the next label at any given point in the CAL algorithm. This refines a related result of El-Yaniv and Wiener (2010). Lemma 8 is also of independent interest in the context of selective prediction (Wiener, 2013; El-Yaniv and Wiener, 2010), as it can be used to improve the known coverage bounds for realizable selective classification.

Lemma 8 *For any* $\delta \in (0,1)$ *and* $m \in \mathbb{N}$*, with probability at least* $1-\delta$*,*

$$\Delta \mathrm{VS}_{\mathcal{F},S_m} \leq \frac{10\hat{n}(S_m)\ln\left(\frac{em}{\hat{n}(S_m)}\right) + 4\ln\left(\frac{2}{\delta}\right)}{m}.$$

Proof The proof is similar to that of a result of El-Yaniv and Wiener (2010), except using a generalization bound based directly on sample compression, rather than the VC dimension. Specifically, let $\mathbb{D} = \{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S}) \times \mathcal{Y} : S \in (\mathcal{X}\times\mathcal{Y})^m\}$, and for each $n \leq m$ and $S \in (\mathcal{X}\times\mathcal{Y})^n$, let $\phi_n(S) = \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S}) \times \mathcal{Y}$. In particular, note that for any $n \geq \hat{n}(S_m)$, any superset S of $\hat{\mathcal{C}}_{S_m}$ of size n contained in S_m has $\phi_n(S) = \mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_m}) \times \mathcal{Y}$, and therefore $S_m \cap \phi_n(S) = \emptyset$ and $\Delta\mathrm{VS}_{\mathcal{F},S_m} = P(\phi_n(S))$. Therefore, Lemma 7 implies that, for each $n \in \{0,\ldots,m\}$, with probability at least $1-\delta/(n+2)^2$, if $\hat{n}(S_m) \leq n$,

$$\Delta\mathrm{VS}_{\mathcal{F},S_m} \leq \frac{1}{m-n}\left(n\ln\left(\frac{em}{n}\right) + \ln\left(\frac{(n+2)^2}{\delta}\right)\right).$$

Furthermore, since $\Delta\mathrm{VS}_{\mathcal{F},S_m} \leq 1$, any $n \geq m/2$ trivially has $\Delta\mathrm{VS}_{\mathcal{F},S_m} \leq 2n/m \leq (2/m)(n\ln(em/n) + \ln((n+2)^2/\delta))$, while any $n \leq m/2$ has $1/(m-n) \leq 2/m$, so that the above is at most

$$\frac{2}{m}\left(n\ln\left(\frac{em}{n}\right) + \ln\left(\frac{(n+2)^2}{\delta}\right)\right).$$

Additionally, $\ln((n+2)^2) \leq 2\ln(2) + 4n \leq 2\ln(2) + 4n\ln(em/n)$, so that the above is at most

$$\frac{2}{m}\left(5n\ln\left(\frac{em}{n}\right) + 2\ln\left(\frac{2}{\delta}\right)\right).$$

By a union bound, this holds for all $n \in \{0,\ldots,m\}$ with probability at least $1-\sum_{n=0}^{m}\delta/(n+2)^2 > 1-\delta$. In particular, since $\hat{n}(S_m)$ is always in $\{0,\ldots,m\}$, this implies the result. ■

Proof of Lemma 6 For any $t \le m$, by definition of $\hat{n}$ (in particular, minimality), *any* set $S \subset S_t$ with $|S| < \hat{n}(S_t)$ necessarily has $\mathrm{VS}_{\mathcal{F},S} \neq \mathrm{VS}_{\mathcal{F},S_t}$. Thus, since CAL maintains that $V_t = \mathrm{VS}_{\mathcal{F},S_t}$, and V_t is precisely the set of classifiers in $\mathcal{F}$ that are correct on the $N(t;S_t)$ points (x_i, y_i) with $i \le t$ for which $\mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{i-1}})}(x_i) = 1$, we must have $N(t;S_t) \ge \hat{n}(S_t)$. We therefore have $\max_{t \le m} \hat{n}(S_t) \le \max_{t \le m} N(t;S_t) = N(m;S_m)$ (by monotonicity of $t \mapsto N(t;S_t)$).

For the upper bound, let δ_i be a sequence of values in $(0,1]$ with $\sum_{i=0}^{\lfloor \log_2(m) \rfloor} \delta_i \le \delta/2$. Lemma 8 implies that, for each i, with probability at least $1 - \delta_i$,

$$\Delta \mathrm{VS}_{\mathcal{F},S_{2^i}} \le 2^{-i} \left(10\hat{n}(S_{2^i}) \ln \left(\frac{e2^i}{\hat{n}(S_{2^i})} \right) + 4 \ln \left(\frac{2}{\delta_i} \right) \right).$$

Thus, by monotonicity of $\Delta \mathrm{VS}_{\mathcal{F},S_t}$ in t, a union bound implies that with probability at least $1 - \delta/2$, for every $i \in \{0, 1, \ldots, \lfloor \log_2(m) \rfloor\}$, every $t \in \{2^i, \ldots, 2^{i+1} - 1\}$ has

$$\Delta \mathrm{VS}_{\mathcal{F},S_t} \le 2^{-i} \left(10\hat{n}(S_{2^i}) \ln \left(\frac{e2^i}{\hat{n}(S_{2^i})} \right) + 4 \ln \left(\frac{2}{\delta_i} \right) \right). \tag{3}$$

Noting that $\left\{ \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{t-1}})}(x_t) - \Delta \mathrm{VS}_{\mathcal{F},S_{t-1}} \right\}_{t=1}^{\infty}$ is a martingale difference sequence with respect to $\{x_t\}_{t=1}^{\infty}$, Bernstein's inequality (for martingales) implies that with probability at least $1 - \delta/2$, if (3) holds for all $i \in \{0, 1, \ldots, \lfloor \log_2(m) \rfloor\}$ and $t \in \{2^i, \ldots, 2^{i+1} - 1\}$, then

$$\begin{aligned} \sum_{t=1}^{m} \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{t-1}})}(x_t) &\le 1 + \sum_{i=0}^{\lfloor \log_2(m) \rfloor} \sum_{t=2^i+1}^{2^{i+1}} \mathbb{1}_{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S_{2^i}})}(x_t) \\ &\le \log_2 \left(\frac{4}{\delta} \right) + 2e \sum_{i=0}^{\lfloor \log_2(m) \rfloor} \left(10\hat{n}(S_{2^i}) \ln \left(\frac{e2^i}{\hat{n}(S_{2^i})} \right) + 4 \ln \left(\frac{2}{\delta_i} \right) \right). \end{aligned}$$

Letting $\delta_i = \frac{\delta}{2 \lfloor \log_2(2m) \rfloor}$, the above is at most

$$\max_{i \in \{0,1,\ldots,\lfloor \log_2(m) \rfloor\}} \left(55\hat{n}(S_{2^i}) \ln \left(\frac{e2^i}{\hat{n}(S_{2^i})} \right) + 24 \ln \left(\frac{4 \log_2(2m)}{\delta} \right) \right) \log_2(2m).$$

■

This also implies distribution-dependent bounds on any confidence bound on the number of queries made by CAL. Specifically, let $\mathcal{B}_N(m, \delta)$ be the smallest nonnegative integer n such that $\mathbb{P}(N(m;S_m) \le n) \ge 1 - \delta$. Then the following result follows immediately from Lemma 6.

Theorem 9 *For any $m \in \mathbb{N}$ and $\delta \in (0,1)$, for any sequence δ_t in $(0,1]$ with $\sum_{i=0}^{\lfloor \log_2(m) \rfloor} \delta_{2^i} \le \delta/2$,*

$$\begin{aligned} \max_{t \le m} \mathcal{B}_{\hat{n}}(t, \delta) &\le \mathcal{B}_N(m, \delta) \\ &\le \max_{t \in \{2^i : i \in \{0,1,\ldots,\lfloor \log_2(m) \rfloor\}\}} \left(55\mathcal{B}_{\hat{n}}(t, \delta_t) \ln \left(\frac{et}{\mathcal{B}_{\hat{n}}(t, \delta_t)} \right) + 24 \ln \left(\frac{8 \log_2(2m)}{\delta} \right) \right) \log_2(2m). \end{aligned}$$

Proof Since Lemma 6 implies every $t \leq m$ has $\hat{n}(S_t) \leq N(m;S_m)$, we have $\mathbb{P}(\hat{n}(S_t) \leq \mathcal{B}_N(m,\delta)) \geq \mathbb{P}(N(m;S_m) \leq \mathcal{B}_N(m,\delta)) \geq 1-\delta$. Since $\mathcal{B}_{\hat{n}}(t,\delta)$ is the smallest $n \in \mathbb{N}$ with $\mathbb{P}(\hat{n}(S_t) \leq n) \geq 1-\delta$, we must therefore have $\mathcal{B}_{\hat{n}}(t,\delta) \leq \mathcal{B}_N(m,\delta)$, from which the left inequality in the claim follows by maximizing over t.

For the second inequality, the upper bound on $N(m;S_m)$ from Lemma 6 implies that, with probability at least $1-\delta/2$, $N(m;S_m)$ is at most

$$\max_{t \in \{2^i : i \in \{0,\ldots,\lfloor \log_2(m) \rfloor\}\}} \left(55\hat{n}(S_t) \ln\left(\frac{et}{\hat{n}(S_t)}\right) + 24 \ln\left(\frac{8\log_2(2m)}{\delta}\right)\right) \log_2(2m).$$

Furthermore, a union bound implies that with probability at least $1 - \sum_{i=0}^{\lfloor \log_2(m) \rfloor} \delta_{2^i} \geq 1-\delta/2$, every $t \in \{2^i : i \in \{0,\ldots,\lfloor \log_2(m) \rfloor\}\}$ has $\hat{n}(S_t) \leq \mathcal{B}_{\hat{n}}(t,\delta_t)$. Since $x \mapsto x\ln(et/x)$ is nondecreasing for $x \in [0,t]$, and $\mathcal{B}_{\hat{n}}(t,\delta_t) \leq t$, combining these two results via a union bound, we have that with probability at least $1-\delta$, $N(m;S_m)$ is at most

$$\max_{t \in \{2^i : i \in \{0,1,\ldots,\lfloor \log_2(m) \rfloor\}\}} \left(55\mathcal{B}_{\hat{n}}(t,\delta_t) \ln\left(\frac{et}{\mathcal{B}_{\hat{n}}(t,\delta_t)}\right) + 24 \ln\left(\frac{8\log_2(2m)}{\delta}\right)\right) \log_2(2m).$$

Letting U_m denote this last quantity, note that since $N(m;S_m)$ is a nonnegative integer, $N(m;S_m) \leq U_m \Rightarrow N(m;S_m) \leq \lfloor U_m \rfloor$, so that $\mathbb{P}(N(m;S_m) \leq \lfloor U_m \rfloor) \geq 1-\delta$. Since $\mathcal{B}_N(m,\delta)$ is the *smallest* nonnegative integer n with $\mathbb{P}(N(m;S_m) \leq n) \geq 1-\delta$, we must have $\mathcal{B}_N(m,\delta) \leq \lfloor U_m \rfloor \leq U_m$. ■

In bounding the label complexity of CAL, we are primarily interested in the size of n sufficient to guarantee low error rate for every classifier in the final V_m set (since $\hat{h}$ is taken to be an arbitrary element of V_m). Specifically, we are interested in the following quantity. For $n \in \mathbb{N}$, define $M(n;S_\infty) = \min\{m \in \mathbb{N} : N(m;S_m) = n\}$ (or $M(n;S_\infty) = \infty$ if $\max_m N(m;S_m) < n$), and for any $\varepsilon, \delta \in (0,1]$, define

$$\Lambda(\varepsilon,\delta) = \min\left\{ n \in \mathbb{N} : \mathbb{P}\left(\sup_{h \in \mathrm{VS}_{\mathcal{F},S_{M(n;S_\infty)}}} \mathrm{er}(h) \leq \varepsilon \right) \geq 1-\delta \right\}.$$

Note that, for any $n \geq \Lambda(\varepsilon,\delta)$, with probability at least $1-\delta$, the classifier $\hat{h}$ produced by CAL(n) has $\mathrm{er}(\hat{h}) \leq \varepsilon$. Furthermore, for any $n < \Lambda(\varepsilon,\delta)$, with probability greater than δ, there exists a choice of $\hat{h}$ in the final step of CAL(n) for which $\mathrm{er}(\hat{h}) > \varepsilon$. Therefore, in a sense, $\Lambda(\varepsilon,\delta)$ represents the label complexity of the general family of CAL strategies (which vary only in how $\hat{h}$ is chosen from the final V_m set). We can also define an analogous quantity for passive learning by empirical risk minimization:

$$M(\varepsilon,\delta) = \min\left\{ m \in \mathbb{N} : \mathbb{P}\left(\sup_{h \in \mathrm{VS}_{\mathcal{F},S_m}} \mathrm{er}(h) \leq \varepsilon \right) \geq 1-\delta \right\}.$$

We typically expect $M(\varepsilon,\delta)$ to be larger than $\Omega(1/\varepsilon)$, and it is known $M(\varepsilon,\delta)$ is always at most $O((1/\varepsilon)(d\log(1/\varepsilon)+\log(1/\delta)))$ (e.g., Vapnik, 1998). We have the following theorem relating these two quantities.

Theorem 10 *There exists a universal constant $c \in (0,\infty)$ such that, $\forall \varepsilon, \delta \in (0,1)$, $\forall \beta \in \left(0, \frac{1-\delta}{\delta}\right)$, for any sequence δ_m in $(0,1]$ with $\sum_{i=0}^{\lfloor \log_2(M(\varepsilon,\delta/2)) \rfloor} \delta_{2^i} \leq \delta/2$,*

$$\max_{m \leq M(\varepsilon, 1-\beta\delta)} \mathcal{B}_{\hat{n}}(m,(1+\beta)\delta) \leq \Lambda(\varepsilon,\delta)$$
$$\leq c\left(\max_{m \leq M(\varepsilon,\delta/2)} \mathcal{B}_{\hat{n}}(m,\delta_m) \ln\left(\frac{em}{\mathcal{B}_{\hat{n}}(m,\delta_m)}\right) + \ln\left(\frac{\log_2(2M(\varepsilon,\delta/2))}{\delta}\right)\right) \log_2(2M(\varepsilon,\delta/2)).$$

Proof By definition of $M(\varepsilon, 1-\beta\delta)$, $\forall m < M(\varepsilon,1-\beta\delta)$, with probability greater than $1-\beta\delta$, $\sup_{h \in \mathrm{VS}_{\mathcal{F},S_m}} \mathrm{er}(h) > \varepsilon$. Furthermore, by definition of $\mathcal{B}_{\hat{n}}(m,(1+\beta)\delta)$, $\forall n < \mathcal{B}_{\hat{n}}(m,(1+\beta)\delta)$, with probability greater than $(1+\beta)\delta$, $\hat{n}(S_m) > n$, which together with Lemma 6 implies $N(m;S_m) > n$, so that $M(n;S_\infty) < m$. Thus, fixing any $m \leq M(\varepsilon,1-\beta\delta)$ and $n < \mathcal{B}_{\hat{n}}(m,(1+\beta)\delta)$, a union bound implies that with probability exceeding δ, $M(n;S_\infty) < m$ and $\sup_{h \in \mathrm{VS}_{\mathcal{F},S_{m-1}}} \mathrm{er}(h) > \varepsilon$. By monotonicity of $t \mapsto \mathrm{VS}_{\mathcal{F},S_t}$, this implies that with probability greater than δ, $\sup_{h\in \mathrm{VS}_{\mathcal{F},S_{M(n;S_\infty)}}} \mathrm{er}(h) > \varepsilon$, so that $\Lambda(\varepsilon,\delta) > n$.

For the upper bound, Lemma 6 and a union bound imply that, with probability at least $1-\delta/2$,

$$N(M(\varepsilon,\delta/2);S_{M(\varepsilon,\delta/2)}) \leq$$
$$c'\left(\max_{m \leq M(\varepsilon,\delta/2)} \mathcal{B}_{\hat{n}}(m,\delta_m) \ln\left(\frac{em}{\mathcal{B}_{\hat{n}}(m,\delta_m)}\right) + \ln\left(\frac{\log_2(2M(\varepsilon,\delta/2))}{\delta}\right)\right) \log_2(2M(\varepsilon,\delta/2)),$$

for a universal constant $c' > 0$. In particular, this implies that for any n at least this large, with probability at least $1-\delta/2$, $M(n+1;S_\infty) \geq M(\varepsilon,\delta/2)$. Furthermore, by definition of $M(\varepsilon,\delta/2)$ and monotonicity of $m \mapsto \sup_{h\in\mathrm{VS}_{\mathcal{F},S_m}} \mathrm{er}(h)$, with probability at least $1-\delta/2$, every $m \geq M(\varepsilon,\delta/2)$ has $\sup_{h \in \mathrm{VS}_{\mathcal{F},S_m}} \mathrm{er}(h) \leq \varepsilon$. By a union bound, with probability at least $1-\delta$, $\sup_{h\in\mathrm{VS}_{\mathcal{F},S_{M(n+1;S_\infty)}}} \mathrm{er}(h) \leq \varepsilon$. This implies $\Lambda(\varepsilon,\delta) \leq n+1$, so that the result holds (for instance, it suffices to take $c = c'+2$). ■

For instance, $\delta_m = \delta/(2\log_2(2M(\varepsilon,\delta/2)))$ might be a natural choice in the above result.

Another implication of these results is a complement to Theorem 4 that was presented in Theorem 5 above.

Proof of Theorem 5 Lemma 29 in Appendix A and monotonicity of $\varepsilon \mapsto \theta(\varepsilon)$ imply that, for $m = \lceil 1/r_0 \rceil$,

$$\mathcal{B}_N(m,\delta) \leq 8 \vee c_0\theta(dr_0/2)\left(d\ln(e\theta(dr_0/2)) + \ln\left(\frac{\log_2(2/r_0)}{\delta}\right)\right)\log_2\left(\frac{2}{r_0}\right)$$
$$\leq (c_0 \vee 8)\theta(dr_0/2)\left(d\ln(e\theta(dr_0/2)) + \ln\left(\frac{\log_2(2/r_0)}{\delta}\right)\right)\log_2\left(\frac{2}{r_0}\right),$$

for a finite universal constant $c_0 > 0$. The result then follows from Theorem 9 and the fact that $\theta(dr_0/2) \leq 2\theta(dr_0)$ (Hanneke, 2014). ■

This also implies the following corollary on the necessary and sufficient conditions for CAL to provide exponential improvements in label complexity when passive learning by empirical risk minimization has $\Omega(1/\varepsilon)$ sample complexity (which is typically the case).[6]

6. All of these equivalences continue to hold even when this $M(\varepsilon,\cdot) = \Omega(1/\varepsilon)$ condition fails, excluding statements 1 and 2, which would then be implied by the others but not vice versa.

Corollary 11 (Characterization of CAL) *If $d < \infty$, and $\exists \delta_0 \in (0,1)$ such that $M(\varepsilon, \delta_0) = \Omega(1/\varepsilon)$, then the following are all equivalent:*

1. $\Lambda(\varepsilon,\delta) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right)$,
2. $\Lambda\left(\varepsilon,\frac{1}{40}\right) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right)$,
3. $\mathcal{B}_{\hat{n}}(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$,
4. $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$,
5. $\theta(r_0) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right)$,
6. $\mathcal{B}_{\Delta}(m,\delta) = O\left(\frac{\text{polylog}(m)}{m}\log\left(\frac{1}{\delta}\right)\right)$,
7. $\mathcal{B}_{\Delta}\left(m,\frac{1}{9}\right) = O\left(\frac{\text{polylog}(m)}{m}\right)$,
8. $\mathcal{B}_{N}(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$,
9. $\mathcal{B}_{N}\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$,

where $\mathcal{F}$ and P are considered constant, so that the big-O hides $(\mathcal{F},P)$-dependent constant factors here (but no factors depending on ε, δ, m, or r_0).[7]

Proof We decompose the proof into a series of implications. Specifically, we show that $3 \Rightarrow 4 \Rightarrow 5 \Rightarrow 8 \Rightarrow 3$, $8 \Rightarrow 9 \Rightarrow 4$, $5 \Rightarrow 1 \Rightarrow 2 \Rightarrow 4$, and $3 \Rightarrow 6 \Rightarrow 7 \Rightarrow 5$. These implications form a strongly connected directed graph, and therefore establish equivalence of the statements.

(3 ⇒ 4) If $\mathcal{B}_{\hat{n}}(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$, then in particular there is some (sufficiently small) constant $\delta_1 \in (0,1/20)$ for which $\mathcal{B}_{\hat{n}}(m,\delta_1) = O(\text{polylog}(m))$, and since $\delta \mapsto \mathcal{B}_{\hat{n}}(m,\delta)$ is nonincreasing, $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) \le \mathcal{B}_{\hat{n}}(m,\delta_1)$, so that $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$ as well.

(4 ⇒ 5) If $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$, then

$$\max_{m \le 1/r_0} \mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O\left(\max_{m\le 1/r_0}\text{polylog}(m)\right) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right).$$

Therefore, Theorem 4 implies

$$\begin{aligned}\theta(r_0) &\le \max\left\{\max_{m\le\lceil 1/r_0\rceil} 16\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right), 512\right\} \\ &\le 528 + 16\max_{m\le 1/r_0}\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right).\end{aligned}$$

7. In fact, we may choose freely whether or not to allow the big-O to hide f^*-dependent constants, or P-dependent constants in general, as long as the *same* interpretation is used for all of these statements. Though validity for each of these interpretations generally does not imply validity for the others, the proof remains valid regardless of which of these interpretations we choose, as long as we stick to the same interpretation throughout the proof.

(5 ⇒ 8) If $\theta(r_0) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right)$, then Lemma 29 in Appendix A implies that $\mathcal{B}_N(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$.

(8 ⇒ 3) If $\mathcal{B}_N(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$, then Theorem 9 implies

$$\mathcal{B}_{\hat{n}}(m,\delta) \leq \mathcal{B}_N(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right).$$

(8 ⇒ 9) If $\mathcal{B}_N(m,\delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$, then for any sufficiently small value $\delta_2 \in (0,1/20)$, $\mathcal{B}_N(m,\delta_2) = O(\text{polylog}(m))$; monotonicity of $\delta \mapsto \mathcal{B}_N(m,\delta)$ further implies $\mathcal{B}_N\left(m,\frac{1}{20}\right) \leq \mathcal{B}_N(m,\delta_2)$, so that $\mathcal{B}_N\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$.

(9 ⇒ 4) When $\mathcal{B}_N\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$, Theorem 9 implies that $\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right) \leq \mathcal{B}_N\left(m,\frac{1}{20}\right) = O(\text{polylog}(m))$.

(5 ⇒ 1) If $\theta(r_0) = O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right)$, then Lemma 30 in Appendix A implies that $\Lambda(\varepsilon,\delta) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right)$.

(1 ⇒ 2) If $\Lambda(\varepsilon,\delta) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right)$, then for any sufficiently small value $\delta_3 \in (0,1/40]$, $\Lambda(\varepsilon,\delta_3) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right)$; furthermore, monotonicity of $\delta \mapsto \Lambda(\varepsilon,\delta)$ implies $\Lambda\left(\varepsilon,\frac{1}{40}\right) \leq \Lambda(\varepsilon,\delta_3)$, so that $\Lambda\left(\varepsilon,\frac{1}{40}\right) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right)$ as well.

(2 ⇒ 4) Let $c \in (0,1]$ and $\varepsilon_0 \in (0,1)$ be constants such that, $\forall \varepsilon \in (0,\varepsilon_0)$, $M(\varepsilon,\delta_0) \geq \frac{c}{\varepsilon}$. For any $\delta \in (0,1/20)$, if $\frac{19}{20}+\delta \leq \delta_0$, then $M\left(\varepsilon,\frac{19}{20}+\delta\right) \geq M(\varepsilon,\delta_0) \geq c/\varepsilon$; otherwise, if $\frac{19}{20}+\delta > \delta_0$, then letting $m = M(\varepsilon,\frac{19}{20}+\delta)$ and $\mathcal{L}_i = \{(x_{m(i-1)+1},y_{m(i-1)+1}),\ldots,(x_{mi},y_{mi})\}$ for $i \in \mathbb{N}$, we have that $\forall k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\left(\sup_{h\in \text{VS}_{\mathcal{F},S_{mk}}} \text{er}(h) > \varepsilon\right) &\leq \mathbb{P}\left(\min_{i\leq k}\sup_{h\in \text{VS}_{\mathcal{F},\mathcal{L}_i}} \text{er}(h) > \varepsilon\right) \\ &= \prod_{i=1}^{k}\mathbb{P}\left(\sup_{h\in \text{VS}_{\mathcal{F},\mathcal{L}_i}} \text{er}(h) > \varepsilon\right) \leq \left(\frac{19}{20}+\delta\right)^k,\end{aligned}$$

so that setting $k = \left\lceil \frac{\ln(1/\delta_0)}{\ln(1/(\frac{19}{20}+\delta))}\right\rceil$ reveals that

$$M(\varepsilon,\delta_0) \leq M\left(\varepsilon,\frac{19}{20}+\delta\right)\left\lceil\frac{\ln(1/\delta_0)}{\ln(1/(\frac{19}{20}+\delta))}\right\rceil. \tag{4}$$

Since $\ln(x) < x-1$ for $x \in (0,1)$, we have $\ln(1/(\frac{19}{20}+\delta)) = -\ln(\frac{19}{20}+\delta) > -(\frac{19}{20}+\delta-1) = \frac{1}{20}-\delta$; together with the fact that $\frac{1}{20}-\delta < 1$, this implies

$$\begin{aligned}\left\lceil\frac{\ln(1/\delta_0)}{\ln(1/(\frac{19}{20}+\delta))}\right\rceil &\leq \left\lceil\frac{\ln(1/\delta_0)}{\frac{1}{20}-\delta}\right\rceil < \frac{\ln(1/\delta_0)}{\frac{1}{20}-\delta}+1 \\ &< \frac{\ln(1/\delta_0)}{\frac{1}{20}-\delta}+\frac{1}{\frac{1}{20}-\delta} = \frac{\ln(e/\delta_0)}{\frac{1}{20}-\delta}.\end{aligned}$$

Plugging this into (4) reveals that

$$M\left(\varepsilon, \frac{19}{20}+\delta\right) \geq \frac{\frac{1}{20}-\delta}{\ln(e/\delta_0)} M(\varepsilon, \delta_0) \geq \frac{c(\frac{1}{20}-\delta)}{\ln(e/\delta_0)} \frac{1}{\varepsilon}.$$

If $\Lambda\left(\varepsilon, \frac{1}{40}\right) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right)$, then Theorem 10 (with $\beta = \frac{1}{20\delta} - 1$ and $\delta = 1/40$) implies

$$\max_{t \leq \frac{c/40}{\ln(e/\delta_0)} \frac{1}{\varepsilon}} \mathcal{B}_{\hat{n}}\left(t, \frac{1}{20}\right) \leq \Lambda\left(\varepsilon, \frac{1}{40}\right) = O\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right).$$

This implies that, $\forall m \in \mathbb{N}$,

$$\begin{aligned} \mathcal{B}_{\hat{n}}\left(m, \frac{1}{20}\right) &\leq \Lambda\left(\frac{c/40}{m\ln(e/\delta_0)}, \frac{1}{40}\right) \\ &= O\left(\text{polylog}\left(\frac{m\ln(e/\delta_0)}{(c/40)}\right)\right) = O\left(\text{polylog}(m)\right). \end{aligned}$$

(3 ⇒ 6) Lemma 8 implies that with probability at least $1-\delta/2$,

$$\Delta \text{VS}_{\mathcal{F}, S_m} \leq \frac{1}{m}\left(10\hat{n}(S_m)\ln\left(\frac{em}{\hat{n}(S_m)}\right) + 4\ln\left(\frac{4}{\delta}\right)\right),$$

while the definition of $\mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right)$ implies that $\hat{n}(S_m) \leq \mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right)$ with probability at least $1-\delta/2$. By a union bound, both of these occur with probability at least $1-\delta$; together with the facts that $x \mapsto x\ln(em/x)$ is nondecreasing on $(0,m]$ and $\mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right) \leq m$, this implies

$$\begin{aligned} \mathcal{B}_{\Delta}(m, \delta) &\leq \frac{1}{m}\left(10\mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right)\ln\left(\frac{em}{\mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right)}\right) + 4\ln\left(\frac{4}{\delta}\right)\right) \\ &= O\left(\frac{1}{m}\left(\mathcal{B}_{\hat{n}}\left(m, \frac{\delta}{2}\right)\log(m) + \log\left(\frac{1}{\delta}\right)\right)\right). \end{aligned}$$

Thus, if $\mathcal{B}_{\hat{n}}(m, \delta) = O\left(\text{polylog}(m)\log\left(\frac{1}{\delta}\right)\right)$, then we have

$$\mathcal{B}_{\Delta}(m, \delta) = O\left(\frac{\text{polylog}(m)}{m}\log\left(\frac{1}{\delta}\right)\right).$$

(6 ⇒ 7) If $\mathcal{B}_{\Delta}(m, \delta) = O\left(\frac{\text{polylog}(m)}{m}\log\left(\frac{1}{\delta}\right)\right)$, then there exists a sufficiently small constant $\delta_4 \in (0, 1/9]$ such that $\mathcal{B}_{\Delta}(m, \delta_4) = O\left(\frac{\text{polylog}(m)}{m}\right)$; in fact, combined with monotonicity of $\delta \mapsto \mathcal{B}_{\Delta}(m, \delta)$, this implies $\mathcal{B}_{\Delta}\left(m, \frac{1}{9}\right) = O\left(\frac{\text{polylog}(m)}{m}\right)$ as well.

(7 ⇒ 5) If $\mathcal{B}_{\Delta}\left(m,\frac{1}{9}\right)=O\left(\frac{\text{polylog}(m)}{m}\right)$, then Lemma 31 in Appendix A implies

$$
\begin{aligned}
\theta(r_0) &\leq \max\left\{\sup_{r\in(r_0,1/2)} \frac{7\mathcal{B}_{\Delta}\left(\lfloor 1/r\rfloor,\frac{1}{9}\right)}{r}, 2\right\} \\
&\leq 2+14\max_{m\leq 1/r_0} m\mathcal{B}_{\Delta}\left(m,\frac{1}{9}\right) \\
&= O\left(\max_{m\leq 1/r_0}\text{polylog}(m)\right)=O\left(\text{polylog}\left(\frac{1}{r_0}\right)\right).
\end{aligned}
$$

■

5. Applications

In this section, we state bounds on the complexity measures studied above, for various hypothesis classes $\mathcal{F}$ and distributions P, which can then be used in conjunction with the above results. In each case, combining the result with theorems above yields a bound on the label complexity of CAL that is smaller than the best known result in the published literature for that problem.

5.1 Linear Separators under Mixtures of Gaussians

The first result, due to El-Yaniv and Wiener (2010), applies to the problem of learning linear separators under a mixture of Gaussians distribution. Specifically, for $k\in\mathbb{N}$, the class of linear separators in $\mathbb{R}^k$ is defined as the set of classifiers $(x_1,\ldots,x_k)\mapsto \text{sign}(b+\sum_{i=1}^k x_i w_i)$, where the values $b,w_1,\ldots,w_k\in\mathbb{R}$ are free parameters specifying the classifier, with $\sum_{i=1}^k w_i^2=1$, and where $\text{sign}(t)=2\mathbb{1}_{[0,\infty)}(t)-1$. In this work, we also include the two constant functions $x\mapsto -1$ and $x\mapsto +1$ as members of the class of linear separators.

Theorem 12 (El-Yaniv and Wiener, 2010, Lemma 32) *For $t,k\in\mathbb{N}$, there is a finite constant $c_{k,t}>0$ such that, for $\mathcal{F}$ the space of linear separators on $\mathbb{R}^k$, and for P with marginal distribution over $\mathcal{X}$ that is a mixture of t multivariate normal distributions with diagonal covariance matrices of full rank, $\forall m\geq 2$,*

$$\mathcal{B}_{\hat{n}}\left(m,\frac{1}{20}\right)\leq c_{k,t}(\log(m))^{k-1}.$$

Combining this result with Theorem 4 implies that there is a constant $c_{k,t}\in(0,\infty)$ such that, for $\mathcal{F}$ and P as in Theorem 12, $\forall r_0\in(0,1/2]$,

$$\theta(r_0)\leq c_{k,t}\left(\log\left(\frac{1}{r_0}\right)\right)^{k-1}.$$

In particular, plugging this into the label complexity bound of Hanneke (2011) for CAL (Lemma 30 of Appendix A) yields the following bound on the label complexity of CAL, which has an improved asymptotic dependence on ε compared to the previous best known result, due to El-Yaniv and Wiener (2012), reducing the exponent on the logarithmic factor from $\Theta(k^2)$ to $\Theta(k)$, and reducing the dependence on δ from $\text{poly}(1/\delta)$ to $\log(1/\delta)$.

Corollary 13 *For $t,k \in \mathbb{N}$, there is a finite constant $c_{k,t} > 0$ such that, for $\mathcal{F}$ the space of linear separators on $\mathbb{R}^k$, and for P with marginal distribution over $\mathcal{X}$ that is a mixture of t multivariate normal distributions with diagonal covariance matrices of full rank, $\forall \varepsilon, \delta \in (0,1/2]$,*

$$\Lambda(\varepsilon,\delta) \le c_{k,t}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^k \log\left(\frac{\log(1/\varepsilon)}{\delta}\right).$$

Corollary 13 is particularly interesting in light of a lower bound of El-Yaniv and Wiener (2012) for this problem, showing that there exists a distribution P of the type described in Corollary 13 for which $\mathcal{B}_N(m,\delta) = \Omega\left((\log(m))^{\frac{k-1}{2}}\right)$.

5.2 Axis-aligned Rectangles under Product Densities

The next result applies to the problem of learning axis-aligned rectangles under product densities over $\mathbb{R}^k$: that is, classifiers $h((x_1',\ldots,x_k')) = 2\prod_{j=1}^k \mathbb{1}_{[a_j,b_j]}(x_j') - 1$, for values $a_1,\ldots,a_k,b_1,\ldots,b_k \in \mathbb{R}$. The result specifically applies to rectangles with a probability at least $\lambda > 0$ of classifying a random point positive. This result represents a refinement of a result of Hanneke (2007b): specifically, reducing a factor of k^2 to a factor of k.

Theorem 14 *For $k,m \in \mathbb{N}$ and $\lambda,\delta \in (0,1)$, for any P with marginal distribution over $\mathcal{X}$ that is a product distribution with marginals having continuous CDFs, and for $\mathcal{F}$ the space of axis-aligned rectangles h on $\mathbb{R}^k$ with $P((x,y): h(x)=1) \ge \lambda$,*

$$\mathcal{B}_{\hat{n}}(m,\delta) \le \frac{8k}{\lambda}\ln\left(\frac{8k}{\delta}\right).$$

Proof The proof is based on a slight refinement of an argument of Hanneke (2007b). For $(X,Y) \sim P$, denote $(X_1,\ldots,X_k) \triangleq X$, let G_i be the CDF of X_i, and define $G(X_1,\ldots,X_k) \triangleq (G_1(X_1),\ldots,G_k(X_k))$. Then the random variable $X' \triangleq (X_1',\ldots,X_k') \triangleq (G_1(X_1),\ldots,G_k(X_k)) = G(X)$ is uniform in $(0,1)^k$; to see this, note that since $X_1,\ldots,X_k$ are independent, so are $G_1(X_1),\ldots,G_k(X_k)$, and that for each $i \le k$, $\forall t \in (0,1)$, $\mathbb{P}(G_i(X_i) \le t) = \sup_{x\in\mathbb{R}:G_i(x)=t}\mathbb{P}(X_i \le x) = \sup_{x\in\mathbb{R}:G_i(x)=t} G_i(x) = t$, where the first equality is by monotonicity and continuity of G_i and the intermediate value theorem (since $\lim_{x\to-\infty} G_i(x) = 0 < t$ and $\lim_{x\to\infty} G_i(x) = 1 > t$), and the second equality is by definition of G_i. Fix any $h \in \mathcal{F}$, let $a_1,\ldots,a_k,b_1,\ldots,b_k \in \mathbb{R}$ be the values such that $h((z_1,\ldots,z_k)) = 2\prod_{i=1}^k \mathbb{1}_{[a_i,b_i]}(z_i) - 1$ for all $(z_1,\ldots,z_k) \in \mathbb{R}^k$, and define $H_h((z_1,\ldots,z_k)) = 2\prod_{i=1}^k \mathbb{1}_{[G_i(a_i),G_i(b_i)]}(z_i) - 1$. Clearly H_h is an axis-aligned rectangle. Furthermore, for every $z \in \mathbb{R}^k$ with $h(z) = +1$, monotonicity of the G_i functions implies $H_h(G(z)) = +1$ as well. Therefore, $\mathbb{P}(H_h(X') = +1) \ge \mathbb{P}(h(X) = +1) \ge \lambda$.

Let $G_i^{-1}(t) = \min\{s : G_i(s) = t\}$ for $t \in (0,1)$, which is well-defined by continuity of G_i and the intermediate value theorem, combined with the facts that $\lim_{z\to\infty} G_i(z) = 1$ and $\lim_{z\to-\infty} G_i(z) = 0$. Let T_i denote the set of discontinuity points of G_i^{-1} in $(0,1)$. Fix any $(z_1,\ldots,z_k) \in \mathbb{R}^k$ with $h((z_1,\ldots,z_k)) = -1$ and $G(z_1,\ldots,z_k) \in (0,1)^k$. In particular, this implies $\exists i \in \{1,\ldots,k\}$ such that $z_i \notin [a_i,b_i]$. For this i, we have $G_i(z_i) \notin (G_i(a_i),G_i(b_i))$ by monotonicity of G_i. Therefore, if $H_h(G(z_1,\ldots,z_k)) = +1$, we must have either $z_i < a_i$ and $G_i(z_i) = G_i(a_i)$, or $z_i > b_i$ and $G_i(z_i) = G_i(b_i)$. In the former case, for any ε with $0 < \varepsilon < 1 - G_i(z_i)$, $G_i^{-1}(G_i(z_i)+\varepsilon) = G_i^{-1}(G_i(a_i)+\varepsilon) > a_i$, while $G_i^{-1}(G_i(z_i)) \le z_i$, and since $z_i < a_i$, we must have $G_i(z_i) \in T_i$. Similarly, in the latter case ($z_i > b_i$ and $G_i(z_i) = G_i(b_i)$), any ε with $0 < \varepsilon < 1 - G_i(z_i)$ has $G_i^{-1}(G_i(b_i)+\varepsilon) = G_i^{-1}(G_i(z_i)+\varepsilon) > z_i$, while $G_i^{-1}(G_i(b_i)) \leq b_i$, and since $z_i > b_i$, we have $G_i(b_i) \in T_i$; since $G_i(z_i) = G_i(b_i)$, this also implies $G_i(z_i) \in T_i$. Thus, any $(z_1,\ldots,z_k) \in \mathbb{R}^k$ with $H_h(G(z_1,\ldots,z_k)) \neq h((z_1,\ldots,z_k))$ must have some $i \in \{1,\ldots,k\}$ with $G_i(z_i) \in T_i$.

For each $i \in \{1,\ldots,k\}$, since G_i is nondecreasing, G_i^{-1} is also nondecreasing, and this implies G_i^{-1} has at most countably many discontinuity points (see e.g., Kolmogorov and Fomin, 1975, Section 31, Theorem 1). Furthermore, for every $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(G_i(X_i) = t) &\leq \mathbb{P}\left(\inf\{x \in \mathbb{R} : G_i(x) = t\} \leq X_i \leq \sup\{x \in \mathbb{R} : G_i(x) = t\}\right)\\ &= G_i(\sup\{x \in \mathbb{R} : G_i(x) = t\}) - G_i(\inf\{x \in \mathbb{R} : G_i(x) = t\}) = t - t = 0,\end{aligned}$$

where the inequality is due to monotonicity of G_i, the first equality is by definition of G_i as the CDF and by continuity of G_i (which implies $\mathbb{P}(X_i < x) = G_i(x)$), and the second equality is due to continuity of G_i. Therefore,

$$\mathbb{P}\left(\exists h \in \mathcal{F} : H_h(G(X)) \neq h(X)\right) \leq \mathbb{P}\left(\exists i \in \{1,\ldots,k\} : G_i(X_i) \in T_i\right) \leq \sum_{i=1}^{k} \sum_{t \in T_i} \mathbb{P}(G_i(X_i) = t) = 0.$$

By a union bound, this implies that with probability 1, for every $h \in \mathcal{F}$, every $(x,y) \in S_m$ has $H_h(G(x)) = h(x)$. In particular, we have that with probability 1, every classification of the sequence $\{x_1,\ldots,x_m\}$ realized by classifiers in $\mathcal{F}$ is also realized as a classification of the i.i.d. Uniform($(0,1)^k$) sequence $\{G(x_1),\ldots,G(x_m)\}$ by the set $\mathcal{F}'$ of axis-aligned rectangles h' with $\mathbb{P}(h'(X') = +1) \geq \lambda$. This implies that $\mathcal{B}_{\hat{n}}(m,\delta) \leq \min\{b \in \mathbb{N} \cup \{0\} : \mathbb{P}(\hat{n}(\mathcal{F}', \{(G(x),y) : (x,y) \in S_m\}) \leq b) \geq 1-\delta\}$ (in fact, one can show they are equal). Therefore, since the right hand side is the value of $\mathcal{B}_{\hat{n}}(m,\delta)$ one would get from the case of P having marginal $P(\cdot \times \mathcal{Y})$ over $\mathcal{X}$ that is Uniform($(0,1)^k$), without loss of generality, it suffices to bound $\mathcal{B}_{\hat{n}}(m,\delta)$ for this special case. Toward this end, for the remainder of this proof, we assume P has marginal $P(\cdot \times \mathcal{Y})$ over $\mathcal{X}$ uniform in $(0,1)^k$.

Let $m \in \mathbb{N}$, and let $\mathcal{U} = \{x_1,\ldots,x_m\}$, the unlabeled portion of the first m data points. Further denote by $\mathcal{U}^+ = \{x_i \in \mathcal{U} : f^*(x_i) = +1\}$, and $\mathcal{U}^- = \mathcal{U} \setminus \mathcal{U}^+$. For each $i \in \mathbb{N}$, express x_i explicitly in vector form as $(x_{i1},\ldots,x_{ik})$. If $\mathcal{U}^+ \neq \emptyset$, for each $j \in \{1,\ldots,k\}$, let $a_j = \min\{x_{ij} : x_i \in \mathcal{U}^+\}$ and $b_j = \max\{x_{ij} : x_i \in \mathcal{U}^+\}$. Denote by $h_{\mathrm{clos}}(x) = 2\mathbb{1}_{\times_{j=1}^{k}[a_j,b_j]}(x) - 1$, the *closure* hypothesis; for completeness, when $\mathcal{U}^+ = \emptyset$, let $h_{\mathrm{clos}}(x) = -1$ for all x.

First, note that if $m < \frac{2e}{\lambda}\left(2k + \ln\left(\frac{2}{\delta}\right)\right)$, the result trivially holds, since $\hat{n}(S_m) \leq m$ always, and $\frac{2e}{\lambda}\left(2k + \ln\left(\frac{2}{\delta}\right)\right) \leq \frac{8k}{\lambda}\ln\left(\frac{8k}{\delta}\right)$. Otherwise, if $m \geq \frac{2e}{\lambda}\left(2k + \ln\left(\frac{2}{\delta}\right)\right)$, a result of Auer and Ortner (2004) implies that, on an event E_{clos} of probability at least $1-\delta/2$, $P((x,y) : h_{\mathrm{clos}}(x) \neq f^*(x)) \leq \lambda/2$. In particular, since $P((x,y) : f^*(x) = +1) \geq \lambda$, on this event we must have $P((x,y) : h_{\mathrm{clos}}(x) = +1) \geq \lambda/2$. Furthermore, this implies $\mathcal{U}^+ \neq \emptyset$ on E_{clos}.

Now fix any $j \in \{1,\ldots,k\}$. Let $x_j^{(aj)}$ denote the value x_{ij} for the point $x_i \in \mathcal{U}$ with largest x_{ij} such that $x_{ij} < a_j$, and for all $j' \neq j$, $x_{ij'} \in [a_{j'},b_{j'}]$; if no such point exists, let $x_j^{(aj)} = 0$. Let $\mathcal{U}^{(aj)} = \{x_i \in \mathcal{U} : x_{ij} < a_j\}$. Let $m^{(aj)} = |\mathcal{U}^{(aj)}|$, and enumerate the points in $\mathcal{U}^{(aj)}$ in decreasing order of x_{ij}, so that $i_1,\ldots,i_{m^{(aj)}}$ are distinct indices such that each $t \in \{1,\ldots,m^{(aj)}\}$ has $x_{i_t} \in \mathcal{U}^{(aj)}$, and each $t \in \{1,\ldots,m^{(aj)}-1\}$ has $x_{i_{t+1}j} \leq x_{i_t j}$. Since $P((x,y) : h_{\mathrm{clos}}(x) = +1) \geq \lambda/2$ on E_{clos}, it must be that the volume of $\times_{j' \neq j}[a_{j'},b_{j'}]$ is at least $\lambda/2$. Therefore, working under the conditional distribution given $\mathcal{U}^+$ and $m^{(aj)}$, on E_{clos}, for each $t \in \{1,\ldots,m^{(aj)}\}$, with conditional probability at least $\lambda/2$, we have $\forall j' \neq j$, $x_{i_t j'} \in [a_{j'},b_{j'}]$. Therefore, the value $t^{(aj)} \triangleq \min\{t : \forall j' \neq j, x_{i_t j'} \in$

$[a_{j'}, b_{j'}]\} \cup \{m^{(aj)}\}$ is bounded by a Geometric random variable with parameter $\lambda/2$. In particular, this implies that with conditional probability at least $1-\frac{\delta}{4k}$, $t^{(aj)} \leq \left\lceil \frac{2}{\lambda}\ln\left(\frac{4k}{\delta}\right)\right\rceil$. Letting $A^{(aj)} = \{x_i \in \mathcal{U} : x_j^{(aj)} \leq x_{ij} < a_j\}$, we note that $|A^{(aj)}| \leq t^{(aj)}$ with probability 1, so that the above reasoning, combined with the law of total probability, implies that there is an event $E^{(aj)}$ of probability at least $1-\frac{\delta}{4k}$ such that, on $E^{(aj)} \cap E_{\text{clos}}$, $|A^{(aj)}| \leq \left\lceil \frac{2}{\lambda}\ln\left(\frac{4k}{\delta}\right)\right\rceil$. For the symmetric case, define $x_j^{(bj)}$ as the value x_{ij} for the point $x_i \in \mathcal{U}$ with smallest x_{ij} such that $x_{ij} > b_j$, and for all $j' \neq j$, $x_{ij'} \in [a_{j'}, b_{j'}]$; if no such point x_i exists, define $x_j^{(bj)} = 1$. Define $A^{(bj)} = \{x_i \in \mathcal{U} : b_j < x_{ij} \leq x_j^{(bj)}\}$. By the same reasoning as above, there is an event $E^{(bj)}$ of probability at least $1-\frac{\delta}{4k}$ such that, on $E^{(bj)} \cap E_{\text{clos}}$, $|A^{(bj)}| \leq \left\lceil \frac{2}{\lambda}\ln\left(\frac{4k}{\delta}\right)\right\rceil$. Applying this to all values of j, and letting $A = \bigcup_{j=1}^{k} A^{(aj)} \cup A^{(bj)}$, we have that on the event $E_{\text{clos}} \cap \bigcap_{j=1}^{k} E^{(aj)} \cap E^{(bj)}$,

$$|A| \leq 2k\left\lceil \frac{2}{\lambda}\ln\left(\frac{4k}{\delta}\right)\right\rceil.$$

Furthermore, a union bound implies that the event $E_{\text{clos}} \cap \bigcap_{j=1}^{k} E^{(aj)} \cap E^{(bj)}$ has probability at least $1-\delta$. For the remainder of the proof, we suppose this event occurs.

Next, let $B = \left\{\operatorname*{argmin}_{x_i \in \mathcal{U}^+} x_{ij} : j \in \{1,\ldots,k\}\right\} \cup \left\{\operatorname*{argmax}_{x_i \in \mathcal{U}^+} x_{ij} : j \in \{1,\ldots,k\}\right\}$, and note that $|B| \leq 2k$. Finally, we conclude the proof by showing that the set $A \cup B$ has the property that $\{h \in \mathcal{F} : \forall x \in A \cup B, h(x) = f^*(x)\} = \mathrm{VS}_{\mathcal{F},S_m}$, which implies $\{(x_i, y_i) : x_i \in A \cup B\}$ is a version space compression set, so that $\hat{n}(S_m) \leq |A \cup B|$, and hence $\mathcal{B}_{\hat{n}}(m,\delta) \leq 2k + 2k\left\lceil \frac{2}{\lambda}\ln\left(\frac{4k}{\delta}\right)\right\rceil \leq \frac{8k}{\lambda}\ln\left(\frac{4k}{\delta}\right)$. To prove that $A \cup B$ has this property, first note that any $h \in \mathcal{F}$ with $h(x_i) = +1$ for all $x_i \in B$, must have $\mathcal{U}^+ \supseteq \{x_i \in \mathcal{U}^+ : h(x_i) = +1\} \supseteq \mathcal{U}^+ \cap \times_{j=1}^{k}[\min_{x_i \in \mathcal{U}^+} x_{ij}, \max_{x_i \in \mathcal{U}^+} x_{ij}] = \mathcal{U}^+$, so that $\{x_i \in \mathcal{U} : h(x_i) = +1\} \supseteq \mathcal{U}^+ = \{x_i \in \mathcal{U} : f^*(x_i) = +1\}$. Next, for any $x_i \in \mathcal{U}^- \setminus (A \cup B)$, $\exists j \in \{1,\ldots,k\} : x_{ij} \notin [a_j, b_j]$, and by definition of A, for this j we must have $x_{ij} \notin [x_j^{(aj)}, x_j^{(bj)}]$. Now fix any $h \in \mathcal{F}$, and express $\{x : h(x) = +1\} = \times_{j'=1}^{k}[a'_{j'}, b'_{j'}]$. If $h(x_{i'}) = +1$ for all $x_{i'} \in B$, then we must have $a'_{j'} \leq a_{j'}$ and $b'_{j'} \geq b_{j'}$ for every $j' \in \{1,\ldots,k\}$. Furthermore, if $h(x_i) = +1$, then we must have $a'_j \leq x_{ij} \leq b'_j$; but then we must have either $a'_j \leq x_{ij} < x_j^{(aj)}$ or $x_j^{(bj)} < x_{ij} \leq b'_j$. In the former case, since $x_{ij} < x_j^{(aj)}$, we must have $x_j^{(aj)} > 0$, so that there exists a point $x_{i'} \in \mathcal{U}$ with $x_{i'j} = x_j^{(aj)}$ and with $x_{i'j'} \in [a_{j'}, b_{j'}]$ for all $j' \neq j$, and furthermore (by definition of A), $x_{i'} \in A$; but since $[a_{j'}, b_{j'}] \subseteq [a'_{j'}, b'_{j'}]$ we also have $x_{i'j'} \in [a'_{j'}, b'_{j'}]$ for all $j' \neq j$, and since $a'_j < x_j^{(aj)} = x_{i'j} < a_j \leq b_j \leq b'_j$, we also have $x_{i'j} \in [a'_j, b'_j]$. Altogether, we must have $h(x_{i'}) = +1$, which proves there exists at least one point in $A \cup B$ classified differently by h and f^*. The case that $x_j^{(bj)} < x_{ij} \leq b'_j$ is symmetric to this one, so that by the same reasoning, this h must disagree with f^* on the classification of some point in $A \cup B$. Therefore, every $h \in \mathcal{F}$ with $h(x) = f^*(x)$ for all $x \in A \cup B$ has $h(x_i) = -1$ for all $x_i \in \mathcal{U}^- \setminus (A \cup B)$. Combined with the above proof that every such h also has $h(x_i) = +1$ for every $x_i \in \mathcal{U}^+$, we have that every such h has $h(x) = f^*(x)$ for every $x \in \mathcal{U}$. ■

One implication of Theorem 14, combined with Theorem 4, is that

$$\theta(r_0) \leq 128\frac{k}{\lambda}\ln(160k)$$

for all $r_0 \geq 0$, for P and $\mathcal{F}$ as in Theorem 14. This has implications, both for the label complexity of CAL (via Lemma 30), and also for the label complexity of noise-robust disagreement-based methods (see Section 6 below). More directly, combining Theorem 14 with Theorem 10 yields the following label complexity bound for CAL, which improves over the best previously published bound on the label complexity of CAL for this problem (due to El-Yaniv and Wiener, 2012), reducing the dependence on k from $\Theta(k^3 \log^2(k))$ to $\Theta(k \log^2(k))$.

Corollary 15 *There exists a finite universal constant $c > 0$ such that, for $k \in \mathbb{N}$ and $\lambda \in (0,1)$, for any P with marginal distribution over $\mathcal{X}$ that is a product distribution with marginals having continuous CDFs, and for $\mathcal{F}$ the space of axis-aligned rectangles h on $\mathbb{R}^k$ with $P((x,y) : h(x) = 1) \geq \lambda$, $\forall \varepsilon, \delta \in (0,1/2)$,*

$$\Lambda(\varepsilon,\delta) \leq c\frac{k}{\lambda}\log\left(\frac{k}{\delta}\log\left(\frac{1}{\varepsilon}\right)\right)\log\left(\frac{k}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{\lambda\log(1/\varepsilon)}{\varepsilon\log(k)} \vee e\right).$$

Proof The result follows by plugging the bound from Theorem 14 into Theorem 10, taking $\delta_m = \delta/(2\log_2(2M(\varepsilon,\delta/2)))$, bounding $M(\varepsilon,\delta/2) \leq \frac{8k}{\varepsilon}\log(\frac{8e}{\varepsilon}) + \frac{8}{\varepsilon}\log(\frac{24}{\delta})$ (Vapnik, 1982; Anthony and Bartlett, 1999), and simplifying the resulting expression. ■

This result is particularly interesting in light of the following lower bound on the label complexities achievable by *any* active learning algorithm.

Theorem 16 *For $k \in \mathbb{N} \setminus \{1\}$ and $\lambda \in (0,1/4]$, letting P_X denote the uniform probability distribution over $(0,1)^k$, for $\mathcal{F}$ the space of axis-aligned rectangles h on $\mathbb{R}^k$ with $P_X(x : h(x) = 1) \geq \lambda$, for any active learning algorithm $\mathcal{A}$, $\forall \delta \in (0,1/2]$, $\forall \varepsilon \in (0,1/(8k))$, there exists a function $f^* \in \mathcal{F}$ such that, if P is the realizable-case distribution having marginal P_X over $\mathcal{X}$ and having target function f^*, if $\mathcal{A}$ is allowed fewer than*

$$\max\left\{k\log\left(\frac{1}{4k\varepsilon}\right), (1-\delta)\left\lfloor\frac{1}{\varepsilon \vee \lambda}\right\rfloor\right\} - 1$$

label requests, then with probability greater than δ, the returned classifier $\hat{h}$ has $\mathrm{er}(\hat{h}) > \varepsilon$.

Proof For any $\varepsilon > 0$, let $\mathcal{M}(\varepsilon)$ denote the maximum number M of classifiers $h_1,\ldots,h_M \in \mathcal{F}$ such that, $\forall i,j \leq M$ with $i \neq j$, $P_X(x : h_i(x) \neq h_j(x)) \geq 2\varepsilon$. Kulkarni, Mitter, and Tsitsiklis (1993) prove that, for any learning algorithm based on binary-valued queries, with a budget smaller than $\log_2((1-\delta)\mathcal{M}(2\varepsilon))$ queries, there exists a target function $f^* \in \mathcal{F}$ such that the classifier $\hat{h}$ produced by the algorithm (when P has marginal P_X over $\mathcal{X}$ and has target function f^*) will have $\mathrm{er}(\hat{h}) > \varepsilon$ with probability greater than δ. In particular, since active learning queries are binary-valued in the binary classification setting, this lower bound applies to active learning algorithms as a special case.

Thus, for the first term in the lower bound, we focus on establishing a lower bound on $\mathcal{M}(2\varepsilon)$ for this problem. First note that $(1-1/k)^k \geq 1/4$, so that $\lambda \leq (1-1/k)^k$. Furthermore, $(1/k)(1-1/k)^{k-1} > 1/(4k)$, so that $\varepsilon < (1/k)(1-1/k)^{k-1}$. Now let

$$\mathcal{F}_{2\varepsilon} = \left\{(x_1,\ldots,x_k) \mapsto 2\prod_{j=1}^{k}\mathbb{1}_{[a_j,b_j]}(x_j) - 1 : \forall j \leq k, b_j = a_j + 1 - 1/k,\right.$$
$$\left. a_j \in \left\{0, \frac{\varepsilon}{(1-1/k)^{k-1}}, \ldots, \left\lfloor\frac{(1-1/k)^{k-1}}{\varepsilon k}\right\rfloor\frac{\varepsilon}{(1-1/k)^{k-1}}\right\}\right\}.$$

Note that $|\mathcal{F}_{2\varepsilon}| = \left(1 + \left\lfloor \frac{(1-1/k)^{k-1}}{\varepsilon k} \right\rfloor\right)^k$. Furthermore, since every $a_j \in [0,1/k]$ in the specification of $\mathcal{F}_{2\varepsilon}$, we have $b_j = a_j + 1 - 1/k \in [0,1]$, which implies $P_X((x_1,\ldots,x_k) : \prod_{j=1}^k \mathbb{1}_{[a_j,b_j]}(x_j) = 1) = (1-1/k)^k \geq \lambda$. Therefore, $\mathcal{F}_{2\varepsilon} \subseteq \mathcal{F}$. Finally, for each $\{(a_j,b_j)\}_{j=1}^k$ and $\{(a'_j,b'_j)\}_{j=1}^k$ specifying distinct classifiers in $\mathcal{F}_{2\varepsilon}$, at least one j has $|a_j - a'_j| \geq \frac{\varepsilon}{(1-1/k)^{k-1}}$. Since all of the elements $h \in \mathcal{F}_{2\varepsilon}$ have $P_X(x : h(x) = +1) = (1-1/k)^k$, we can note that

$$\begin{aligned}
&P_X\left((x_1,\ldots,x_k) : \prod_{i=1}^k \mathbb{1}_{[a_i,b_i]}(x_i) \neq \prod_{i=1}^k \mathbb{1}_{[a'_i,b'_i]}(x_i)\right) \\
&= 2(1-1/k)^k - 2P_X\left((\times_{i=1}^k [a_i,b_i]) \cap (\times_{i=1}^k [a'_i,b'_i])\right) \\
&= 2(1-1/k)^k - 2P_X\left(\times_{i=1}^k [\max\{a_i,a'_i\}, \min\{b_i,b'_i\}]\right) \\
&= 2(1-1/k)^k - 2\prod_{i=1}^k (\min\{b_i,b'_i\} - \max\{a_i,a'_i\}).
\end{aligned}$$

Thus, since

$$\begin{aligned}
&\prod_{i=1}^k (\min\{b_i,b'_i\} - \max\{a_i,a'_i\}) \\
&\leq (\min\{b_j,b'_j\} - \max\{a_j,a'_j\}) \prod_{i \neq j} (b_i - a_i) = (1-1/k)^{k-1}(\min\{b_j,b'_j\} - \max\{a_j,a'_j\}) \\
&= (1-1/k)^{k-1}(\min\{a_j,a'_j\} - \max\{a_j,a'_j\} + (1-1/k)) = (1-1/k)^{k-1}(1-1/k - |a_j - a'_j|) \\
&\leq (1-1/k)^{k-1}(1-1/k - \frac{\varepsilon}{(1-1/k)^{k-1}}) = (1-1/k)^k - \varepsilon,
\end{aligned}$$

we have

$$P_X((x_1,\ldots,x_k) : \prod_{i=1}^k \mathbb{1}_{[a_i,b_i]}(x_i) \neq \prod_{i=1}^k \mathbb{1}_{[a'_i,b'_i]}(x_i)) \geq 2(1-1/k)^k - 2((1-1/k)^k - \varepsilon) = 2\varepsilon.$$

Thus, $\mathcal{M}(2\varepsilon) \geq \left(1 + \left\lfloor \frac{(1-1/k)^{k-1}}{\varepsilon k} \right\rfloor\right)^k$. Finally, note that for $\delta \in (0,1/2]$, this implies

$$\log_2((1-\delta)\mathcal{M}(2\varepsilon)) \geq k\log_2\left(\frac{(1-1/k)^{k-1}}{\varepsilon k}\right) - 1 \geq k\log_2\left(\frac{1}{4k\varepsilon}\right) - 1.$$

Together with the aforementioned lower bound of Kulkarni, Mitter, and Tsitsiklis (1993), this establishes the first term in the lower bound.

To prove the second term, we use of a technique of Hanneke (2007b). Specifically, fix any finite set $H \subseteq \mathcal{F}$ with $\min_{h,g \in H} P_X(x : h(x) \neq g(x)) \geq 2\varepsilon$, let

$$\mathrm{XPTD}(f,H,\mathcal{U},\delta) = \min\{t \in \mathbb{N} : \exists R \subseteq \mathcal{U} : |R| \leq t, |\{h \in H : \forall x \in R, h(x) = f(x)\}| \leq \delta|H| + 1\} \cup \{\infty\},$$

for any classifier f and $\mathcal{U} \in \bigcup_m \mathcal{X}^m$, and let $\mathrm{XPTD}(H,P_X,\delta)$ denote the smallest $t \in \mathbb{N}$ such that every classifier f has $\lim_{m\to\infty} \mathbb{P}_{\mathcal{U}\sim P_X^m}(\mathrm{XPTD}(f,H,\mathcal{U},\delta) > t) = 0$. Then Hanneke (2007b) proves that there exists a choice of target function $f^* \in \mathcal{F}$ for the distribution P such that, if $\mathcal{A}$ is allowed fewer than $\mathrm{XPTD}(H,P_X,\delta)$ label requests, then with probability greater than δ, the returned classifier $\hat{h}$ has $\mathrm{er}(\hat{h}) > \varepsilon$. For the particular problem studied here, let H be the set of classifiers $h_i(x) = 2\mathbb{1}_{[(i-1)(\varepsilon\vee\lambda), i(\varepsilon\vee\lambda)]\times[0,1]^{k-1}}(x) - 1$, for $i \in \left\{1, \ldots, \left\lfloor \frac{1}{\varepsilon\vee\lambda} \right\rfloor\right\}$. Note that each $h_i \in H$ has $P_X(x : h_i(x) = +1) = P_X((x_1, \ldots, x_k) : x_1 \in [(i-1)(\varepsilon\vee\lambda), i(\varepsilon\vee\lambda)]) = \varepsilon\vee\lambda \geq \lambda$, so that $H \subseteq \mathcal{F}$. Furthermore, for any $h_i, h_j \in H$ with $i \neq j$, $P_X(x : h_i(x) \neq h_j(x)) \geq P_X((x_1, \ldots, x_k) : x_1 \in ((i-1)(\varepsilon\vee\lambda), i(\varepsilon\vee\lambda)) \cup ((j-1)(\varepsilon\vee\lambda), j(\varepsilon\vee\lambda))) = 2(\varepsilon\vee\lambda) \geq 2\varepsilon$. Also, let $R \subseteq (0,1)^k$ be any finite set with no points $(x_1, \ldots, x_k) \in R$ such that $x_1 \in \left\{i(\varepsilon\vee\lambda) : i \in \left\{1, \ldots, \left\lfloor \frac{1}{\varepsilon\vee\lambda} \right\rfloor - 1\right\}\right\}$; note that every $x \in R$ has exactly one $h_i \in H$ with $h_i(x) = +1$. Thus, for the classifier f with $f(x) = -1$ for all $x \in \mathcal{X}$, $|\{h \in H : \forall x \in R, h(x) = f(x)\}| \geq |H| - |R|$. Thus, for any set $\mathcal{U} \subseteq (0,1)^k$ with no points $(x_1, \ldots, x_k) \in \mathcal{U}$ having $x_1 \in \left\{i(\varepsilon\vee\lambda) : i \in \left\{1, \ldots, \left\lfloor \frac{1}{\varepsilon\vee\lambda} \right\rfloor - 1\right\}\right\}$, we have $\mathrm{XPTD}(f, H, \mathcal{U}, \delta) \geq (1-\delta)|H| - 1$. Since, for all $m \in \mathbb{N}$, the probability that $\mathcal{U} \sim P_X^m$ contains a point $(x_1, \ldots, x_k)$ with $x_1 \in \left\{i(\varepsilon\vee\lambda) : i \in \left\{1, \ldots, \left\lfloor \frac{1}{\varepsilon\vee\lambda} \right\rfloor - 1\right\}\right\}$ is zero, we have that $\mathbb{P}_{\mathcal{U}\sim P_X^m}(\mathrm{XPTD}(f, H, \mathcal{U}, \delta) \geq (1-\delta)|H| - 1) = 1$. This implies $\mathrm{XPTD}(H, P_X, \delta) \geq (1-\delta)|H| - 1 = (1-\delta)\left\lfloor \frac{1}{\varepsilon\vee\lambda} \right\rfloor - 1$. Combining this with the lower bound of Hanneke (2007b) implies the result. ■

Together, Corollary 15 and Theorem 16 imply that, for $\lambda \in (0, 1/4]$ bounded away from 0, the label complexity of CAL is within logarithmic factors of the minimax optimal label complexity.

6. New Label Complexity Bounds for Agnostic Active Learning

In this section we present new bounds on the label complexity of noise-robust active learning algorithms, expressed in terms of $\mathcal{B}_{\hat{n}}(m, \delta)$. These bounds yield new exponential label complexity speedup results for agnostic active learning (for the low accuracy regime) of linear classifiers under a fixed mixture of Gaussians. Analogous results also hold for the problem of learning axis-aligned rectangles under a product density.

Specifically, in the *agnostic* setting studied in this section, we no longer assume $\exists f^* \in \mathcal{F}$ with $\mathbb{P}(Y = f^*(x)|X) = 1$ for $(X,Y) \sim P$, but rather allow that P is *any* probability measure over $\mathcal{X} \times \mathcal{Y}$. In this setting, we let $f^* : \mathcal{X} \to \mathcal{Y}$ denote a classifier such that $\mathrm{er}(f^*) = \inf_{h\in\mathcal{F}} \mathrm{er}(h)$ and $\inf_{h\in\mathcal{F}} P((x,y) : h(x) \neq f^*(x)) = 0$, which is guaranteed to exist by topological considerations (see Hanneke, 2012, Section 6.1);[8] for simplicity, when $\exists f \in \mathcal{F}$ with $\mathrm{er}(f) = \inf_{h\in\mathcal{F}} \mathrm{er}(h)$, we take f^* to be an element of $\mathcal{F}$. We call f^* the *infimal* hypothesis (of $\mathcal{F}$, w.r.t. P) and note that $\mathrm{er}(f^*)$ is sometimes called the *noise rate of* $\mathcal{F}$ (e.g., Balcan, Beygelzimer, and Langford, 2006). The introduction of the infimal hypothesis f^* allows for natural generalizations of some of the key definitions of Section 2 that facilitate analysis in the agnostic setting.

Definition 17 (Agnostic Version Space) *Let f^* be the infimal hypothesis of $\mathcal{F}$ w.r.t. P. The* agnostic version space *of a sample S is*

$$\mathrm{VS}_{\mathcal{F},S,f^*} \triangleq \{h \in \mathcal{F} : \forall (x,y) \in S, h(x) = f^*(x)\}.$$

Definition 18 (Agnostic Version Space Compression Set Size) *Letting $\hat{C}_{S,f^*}$ denote a smallest subset of S satisfying $\mathrm{VS}_{\mathcal{F},\hat{C}_{S,f^*},f^*} = \mathrm{VS}_{\mathcal{F},S,f^*}$, the* agnostic version space compression set size *is*

$$\hat{n}(\mathcal{F}, S, f^*) \triangleq |\hat{C}_{S,f^*}|.$$

8. In the agnostic setting, there are typically many valid choices of the function f^* satisfying these conditions. The results below hold for *any* such choice of f^*.

We also extend the definition of the version space compression set minimal *bound* (see Definition 2) to the agnostic setting, defining

$$\mathcal{B}_{\hat{n}}(m,\delta) \triangleq \min\{b \in \mathbb{N} \cup \{0\} : \mathbb{P}(\hat{n}(\mathcal{F},S,f^*) \leq b) \geq 1-\delta\}.$$

For general P in the agnostic setting, define the disagreement coefficient as before, except now with respect to the infimal hypothesis:

$$\theta(r_0) \triangleq \sup_{r>r_0} \frac{\Delta \mathrm{B}(f^*,r)}{r} \vee 1.$$

One can easily verify that these definitions are equal to those given above in the special case that P satisfies the realizable-case assumptions ($f^* \in \mathcal{F}$ and $\mathbb{P}(Y = f^*(X)|X) = 1$ for $(X,Y) \sim P$).

We begin with the following extension of Theorem 4.

Lemma 19 *For general (agnostic) P, for any $r_0 \in (0,1)$,*

$$\theta(r_0) \leq \max\left\{\max_{r\in(r_0,1)} 16\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right), 512\right\}.$$

Proof First note that $\theta(r_0)$ and $\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right)$ depend on P only via f^* and the marginal $P(\cdot\times\mathcal{Y})$ of P over $\mathcal{X}$ (in both the realizable case and agnostic case). Define a distribution P' with marginal $P'(\cdot\times\mathcal{Y}) = P(\cdot\times\mathcal{Y})$ over $\mathcal{X}$, and with $\mathbb{P}(Y = f^*(x)|X = x) = 1$ for all $x \in \mathcal{X}$, where $(X,Y) \sim P'$. In particular, in the special case that $f^* \in \mathcal{F}$ in the agnostic case, we have that P' is a distribution in the realizable case, with identical values of $\theta(r_0)$ and $\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right)$ as P, so that Theorem 4 (applied to P') implies the result. On the other hand, when P is a distribution with $f^* \notin \mathcal{F}$, let $\theta'(r_0)$ denote the disagreement coefficient of $\mathcal{F}\cup\{f^*\}$ with respect to P' (or equivalently P), and for $m \in \mathbb{N}$, let $\mathcal{B}'_{\hat{n}}(m,1/20) \triangleq \min\{b \in \mathbb{N}\cup\{0\} : \mathbb{P}(\hat{n}(\mathcal{F}\cup\{f^*\},S_m,f^*) \leq b) \geq 19/20\}$. In particular, since $\mathcal{F} \subseteq \mathcal{F}\cup\{f^*\}$, we have $\theta(r_0) \leq \theta'(r_0)$, and since P' is a realizable-case distribution with respect to the hypothesis class $\mathcal{F}\cup\{f^*\}$, Theorem 4 (applied to P' and $\mathcal{F}\cup\{f^*\}$) implies

$$\theta'(r_0) \leq \max\left\{\max_{r\in(r_0,1)} 16\mathcal{B}'_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right), 512\right\}.$$

Finally, note that for any $m \in \mathbb{N}$ and sets $C,S \in (\mathcal{X}\times\mathcal{Y})^m$, $\mathrm{VS}_{\mathcal{F}\cup\{f^*\},C,f^*} = \mathrm{VS}_{\mathcal{F},C,f^*}\cup\{f^*\}$ and $\mathrm{VS}_{\mathcal{F}\cup\{f^*\},S,f^*} = \mathrm{VS}_{\mathcal{F},S,f^*}\cup\{f^*\}$, so that $\mathrm{VS}_{\mathcal{F}\cup\{f^*\},C,f^*} = \mathrm{VS}_{\mathcal{F}\cup\{f^*\},S,f^*}$ if and only if $\mathrm{VS}_{\mathcal{F},C,f^*} = \mathrm{VS}_{\mathcal{F},S,f^*}$. Thus, $\hat{n}(\mathcal{F}\cup\{f^*\},S_m,f^*) = \hat{n}(\mathcal{F},S_m,f^*)$, so that $\mathcal{B}'_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right) = \mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right)$, which implies the result. ∎

6.1 Label complexity bound for agnostic active learning

A^2 (*Agnostic Active*) was the first general-purpose agnostic active learning algorithm with proven improvement in error guarantees compared to passive learning. The original work of Balcan, Beygelzimer, and Langford (2006), which first introduced this algorithm, also provided specialized proofs that the algorithm achieves an exponential label complexity speedup (for the low accuracy regime) compared to passive learning for a few simple cases, including: threshold functions, and homogeneous linear separators under a uniform distribution over the sphere. Additionally, Hanneke (2007a) provided a general bound on the label complexity of A^2, expressed in terms of the disagreement coefficient, so that any bound on the disagreement coefficient translates into a bound on the label complexity of agnostic active learning with A^2. Inspired by the A^2 algorithm, other noise-robust active learning algorithms have since been proposed, with improved label complexity bounds compared to those proven by Hanneke (2007a) for A^2, while still expressed in terms of the disagreement coefficient (see e.g., Dasgupta, Hsu, and Monteleoni, 2007; Hanneke, 2014). As an example of such results, the following result was proven by Dasgupta, Hsu, and Monteleoni (2007).

Theorem 20 (Dasgupta, Hsu, and Monteleoni, 2007) *There exists a finite universal constant $c > 0$ such that, for any $\varepsilon, \delta \in (0, 1/2)$, using hypothesis class $\mathcal{F}$, and given the input δ and a budget n on the number of label requests, the active learning algorithm of Dasgupta, Hsu, and Monteleoni (2007) requests at most n labels,*[9] *and if*

$$n \geq c\theta(\mathrm{er}(f^*) + \varepsilon) \left(\frac{\mathrm{er}(f^*)^2}{\varepsilon^2} + 1 \right) \left(d \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right) \log\left(\frac{1}{\varepsilon}\right),$$

then with probability at least $1 - \delta$, the classifier $\hat{f} \in \mathcal{F}$ it produces satisfies

$$\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon.$$

Combined with the results above, this implies the following theorem.

Theorem 21 *There exists a finite universal constant $c > 0$ such that, for any $\varepsilon, \delta \in (0, 1/2)$, using hypothesis class $\mathcal{F}$, and given the input δ and a budget n on the number of label requests, the active learning algorithm of Dasgupta, Hsu, and Monteleoni (2007) requests at most n labels, and if*

$$n \geq c \left(\max_{r > \mathrm{er}(f^*) + \varepsilon} \mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20} \right) + 1 \right) \left(\frac{\mathrm{er}(f^*)^2}{\varepsilon^2} + 1 \right) \left(d \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right) \log\left(\frac{1}{\varepsilon}\right),$$

then with probability at least $1 - \delta$, the classifier $\hat{f} \in \mathcal{F}$ it produces satisfies

$$\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon.$$

Proof By Lemma 19,

$$\begin{aligned} \theta(\mathrm{er}(f^*) + \varepsilon) &\leq \max\left\{ \max_{r \in (\mathrm{er}(f^*) + \varepsilon, 1)} 16 \mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20} \right), 512 \right\} \\ &\leq 512 \left(\max_{r > \mathrm{er}(f^*) + \varepsilon} \mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20} \right) + 1 \right). \end{aligned}$$

Plugging this into Theorem 20 yields the result. ∎

9. This result applies to a slightly modified variant of the algorithm of Dasgupta, Hsu, and Monteleoni (2007), studied by Hanneke (2011), which terminates after a given number of label requests, rather than after a given number of unlabeled samples. The same is true of Theorem 21 and Corollary 22.

Interestingly, from the perspective of bounding the label complexity of agnostic active learning in general, the result in Theorem 21 sometimes improves over a related bound proven by Hanneke (2007b) (for a different algorithm). Specifically, compared to the result of Hanneke (2007b), this result maintains an interesting dependence on f^*, whereas the bound of Hanneke (2007b) effectively replaces the factor $\mathcal{B}_{\hat{n}}(\lceil 1/r \rceil, 1/20)$ with the maximum of this quantity over the choice of f^*.[10] Also, while the result of Hanneke (2007b) is proven for an algorithm that requires explicit access to a value $\eta \approx \mathrm{er}(f^*)$ to obtain the stated label complexity, the label complexity in Theorem 21 is achieved by the algorithm of Dasgupta, Hsu, and Monteleoni (2007), which requires no such extra parameters.

As an application of Theorem 21, we have the following corollary.

Corollary 22 *For $t, k \in \mathbb{N}$ and $c \in (0, \infty)$, there exists a finite constant $c_{k,t,c} > 0$ such that, for $\mathcal{F}$ the class of linear separators on $\mathbb{R}^k$, and for P with marginal distribution over $\mathcal{X}$ that is a mixture of t multivariate normal distributions with diagonal covariance matrices of full rank, for any $\varepsilon, \delta \in (0, 1/2)$ with $\varepsilon \geq \frac{\mathrm{er}(f^*)}{c}$, using hypothesis class $\mathcal{F}$, and given the input δ and a budget n on the number of label requests, the active learning algorithm of Dasgupta, Hsu, and Monteleoni (2007) requests at most n labels, and if*

$$n \geq c_{k,t,c} \left(\log \left(\frac{1}{\varepsilon} \right) \right)^{k+1} \log \left(\frac{1}{\delta} \right),$$

then with probability at least $1 - \delta$, the classifier $\hat{f} \in \mathcal{F}$ it produces satisfies $\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon$.

Proof Let $\mathcal{F}$ and P be as described above. First, we argue that $f^* \in \mathcal{F}$. Fix any classifier f with $\inf_{h \in \mathcal{F}} P((x,y) : h(x) \neq f(x)) = 0$. There must exist a sequence $\{(b^{(t)}, w_1^{(t)}, \ldots, w_k^{(t)})\}_{k=1}^{\infty}$ in $\mathbb{R}^{k+1}$ with $\sum_{i=1}^k (w_i^{(t)})^2 = 1$ for all t, s.t. $P\left((x_1, \ldots, x_k, y) : \mathrm{sign}\left(b^{(t)} + \sum_{i=1}^k x_i w_i^{(t)}\right) \neq f(x_1, \ldots, x_k)\right) \to 0$. If $\limsup_{t \to \infty} b^{(t)} = \infty$, then $\exists t_j \to \infty$ with $b^{(t_j)} \to \infty$, and since every $(x_1, \ldots, x_k) \in \mathbb{R}^k$ has $\sum_{i=1}^k x_i w_i^{(t)} \geq -\|x\|$, we have that $b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)} \to \infty$, which implies $\mathrm{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \to 1$ for all $(x_1, \ldots, x_k) \in \mathbb{R}^k$. Similarly, if $\liminf_{t \to \infty} b^{(t)} = -\infty$, then $\exists t_j \to \infty$ with $\mathrm{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \to -1$ for all $(x_1, \ldots, x_k) \in \mathbb{R}^k$. Otherwise, if $\limsup_{t \to \infty} b^{(t)} < \infty$ and $\liminf_{t \to \infty} b^{(t)} > -\infty$, then the sequence $\{(b^{(t)}, w_1^{(t)}, \ldots, w_k^{(t)})\}_{t=1}^{\infty}$ is *bounded* in $\mathbb{R}^{k+1}$. Therefore, the Bolzano-Weierstrass Theorem implies it contains a convergent subsequence: that is, $\exists t_j \to \infty$ s.t. $(b^{(t_j)}, w_1^{(t_j)}, \ldots, w_k^{(t_j)})$ converges. Furthermore, since $\{w \in \mathbb{R}^k : \|w\| = 1\}$ is closed, and $\{b^{(t)} : t \in \mathbb{N}\} \subseteq [\inf_t b^{(t)}, \sup_t b^{(t)}]$, which is a closed subset of $\mathbb{R}$, $\exists (b, w_1, \ldots, w_k) \in \mathbb{R}^{k+1}$ with $\sum_{i=1}^k w_i^2 = 1$ such that $(b^{(t_j)}, w_1^{(t_j)}, \ldots, w_k^{(t_j)}) \to (b, w_1, \ldots, w_k)$. Continuity of linear functions implies, $\forall (x_1, \ldots, x_k) \in \mathbb{R}^k$, $b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)} \to b + \sum_{i=1}^k x_i w_i$. Therefore, every $(x_1, \ldots, x_k) \in \mathbb{R}^k$ with $b + \sum_{i=1}^k x_i w_i > 0$ has $\mathrm{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \to 1$, and every $(x_1, \ldots, x_k) \in \mathbb{R}^k$ with $b + \sum_{i=1}^k x_i w_i < 0$ has $\mathrm{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \to -1$. Since $P\left((x_1, \ldots, x_k, y) : b + \sum_{i=1}^k x_i w_i = 0\right) = 0$, this implies $(x_1, \ldots, x_k) \mapsto \mathrm{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right)$ converges to $(x_1, \ldots, x_k) \mapsto \mathrm{sign}\left(b + \sum_{i=1}^k x_i w_i\right)$ almost surely [P].

10. There are a few other differences, which are usually minor. For instance, the bound of Hanneke (2007b) uses $r \approx \mathrm{er}(f^*) + \varepsilon$ rather than maximizing over $r > \mathrm{er}(f^*) + \varepsilon$. That result additionally replaces "$1/20$" with a value $\delta' \approx \delta/n$.

Thus, in each case, $\exists t_j \to \infty$ and $h \in \mathcal{F}$ s.t. $(x_1,\ldots,x_k) \mapsto \operatorname{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right)$ converges to h a.s. $[P]$. Since convergence almost surely implies convergence in probability, we have $P\left((x_1,\ldots,x_k,y) : \operatorname{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \neq h(x_1,\ldots,x_k)\right) \to 0$. Furthermore, by assumption, $P\left((x_1,\ldots,x_k,y) : \operatorname{sign}\left(b^{(t_j)} + \sum_{i=1}^k x_i w_i^{(t_j)}\right) \neq f(x_1,\ldots,x_k)\right) \to 0$ as well. Thus, a union bound implies $P((x,y) : h(x) \neq f(x)) = 0$. In particular, we have that for any f with $\inf_{g\in\mathcal{F}} P((x,y) : g(x) \neq f(x)) = 0$ and $\operatorname{er}(f) = \inf_{g\in\mathcal{F}} \operatorname{er}(g)$, $\exists h \in \mathcal{F}$ with $P((x,y) : f(x) \neq h(x)) = 0$, and hence $\operatorname{er}(h) = \inf_{g\in\mathcal{F}} \operatorname{er}(g)$. Thus, we may assume $f^* \in \mathcal{F}$ in this setting.

Therefore, in this scenario, Theorem 12 implies

$$\max_{r>\operatorname{er}(f^*)+\varepsilon} \mathcal{B}_{\hat{n}}\left(\left\lceil \frac{1}{r} \right\rceil, \frac{1}{20}\right) + 1 \leq c_{k,t}^{(1)} \left(\log\left(\frac{2}{\operatorname{er}(f^*)+\varepsilon}\right)\right)^{k-1},$$

for an appropriate (k,t)-dependent constant $c_{k,t}^{(1)} \in (0,\infty)$. Plugging this into Theorem 21, and recalling that the VC dimension of the class of linear classifiers in $\mathbb{R}^k$ is $k+1$ (see e.g., Anthony and Bartlett, 1999), we get a bound on the number of label requests of

$$\begin{aligned} & c_{k,t}^{(2)} \left(\log\left(\frac{2}{\operatorname{er}(f^*)+\varepsilon}\right)\right)^{k-1} \left(\frac{\operatorname{er}(f^*)^2}{\varepsilon^2} + 1\right) \left(k \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right) \log\left(\frac{1}{\varepsilon}\right) \\ & \leq c_{k,t}^{(3)} \left(\log\left(\frac{1}{\varepsilon}\right)\right)^{k+1} \left(\frac{\operatorname{er}(f^*)^2}{\varepsilon^2} + 1\right) \left(k + \log\left(\frac{1}{\delta}\right)\right), \end{aligned}$$

for appropriate (k,t)-dependent constants $c_{k,t}^{(2)}, c_{k,t}^{(3)} \in (0,\infty)$. Since (by assumption) $\varepsilon \geq \frac{\operatorname{er}(f^*)}{c}$, this is at most

$$c_{k,t,c}^{(4)} \left(\log\left(\frac{1}{\varepsilon}\right)\right)^{k+1} \left(k + \log\left(\frac{1}{\delta}\right)\right) \leq c_{k,t,c}^{(5)} \left(\log\left(\frac{1}{\varepsilon}\right)\right)^{k+1} \log\left(\frac{1}{\delta}\right),$$

for appropriate (k,t,c)-dependent constants $c_{k,t,c}^{(4)}, c_{k,t,c}^{(5)} \in (0,\infty)$. Thus, taking $c_{k,t,c} = c_{k,t,c}^{(5)}$ establishes the result. ■

An analogous result can be shown for the problem of learning axis-aligned rectangles via Theorem 14.

6.2 Label complexity bound under Mammen-Tsybakov noise

Since the original work on agnostic active learning discussed above, there have been several other analyses, expressing the noise conditions in terms of quantities other than the noise rate $\operatorname{er}(f^*)$. Specifically, the following condition of Mammen and Tsybakov (1999) has been studied for several algorithms (see e.g., Balcan, Broder, and Zhang, 2007; Hanneke, 2011; Koltchinskii, 2010; Hanneke, 2012; Hanneke and Yang, 2012; Hanneke, 2014; Beygelzimer, Hsu, Langford, and Zhang, 2010; Hsu, 2010).

Condition 23 (Mammen and Tsybakov, 1999) *For some $a \in [1,\infty)$ and $\alpha \in [0,1]$, for every $f \in \mathcal{F}$,*

$$\Pr(f(X) \neq f^*(X)) \leq a(\operatorname{er}(f) - \operatorname{er}(f^*))^{\alpha}.$$

In particular, for a variant of A^2 known as RobustCAL$_\delta$, studied by Hanneke (2012, 2014) and Hanneke and Yang (2012), the following result is known (due to Hanneke and Yang, 2012).

Theorem 24 (Hanneke and Yang, 2012) *There exists a finite universal constant $c > 0$ such that, for any $\varepsilon, \delta \in (0, 1/2)$, for any $n, u \in \mathbb{N}$, given the arguments n and u, the RobustCAL$_\delta$ algorithm requests at most n labels, and if u is sufficiently large, and*

$$n \geq ca^2\theta(a\varepsilon^\alpha)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\log\left(e\theta\left(a\varepsilon^\alpha\right)\right)+\log\left(\frac{\log(1/\varepsilon)}{\delta}\right)\right)\log\left(\frac{1}{\varepsilon}\right),$$

for a and α as in Condition 23, then with probability at least $1-\delta$, the classifier $\hat{f} \in \mathcal{F}$ it returns satisfies $\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon$.

Combined with Theorem 4, this implies the following theorem.

Theorem 25 *There exists a finite universal constant $c > 0$ such that, for any $\varepsilon, \delta \in (0, 1/2)$, for any $n, u \in \mathbb{N}$, given the arguments n and u, the RobustCAL$_\delta$ algorithm requests at most n labels, and if u is sufficiently large, and*

$$n \geq ca^2\left(\max_{r > a\varepsilon^\alpha}\mathcal{B}_{\hat{n}}\left(\left\lceil\frac{1}{r}\right\rceil, \frac{1}{20}\right)+1\right)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{1}{\varepsilon}\right),$$

for a and α as in Condition 23, then with probability at least $1-\delta$, the classifier $\hat{f} \in \mathcal{F}$ it returns satisfies $\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon$.

In particular, reasoning as in Corollary 22 above, Theorem 25 implies the following corollary.

Corollary 26 *For $t, k \in \mathbb{N}$ and $a \in [1, \infty)$, there exists a finite constant $c_{k,t,a} > 0$ such that, for $\mathcal{F}$ the class of linear separators on $\mathbb{R}^k$, and for P satisfying Condition 23 with $\alpha = 1$ and the given value of a, and with marginal distribution over $\mathcal{X}$ that is a mixture of t multivariate normal distributions with diagonal covariance matrices of full rank, for any $\varepsilon, \delta \in (0, 1/2)$, for any $n, u \in \mathbb{N}$, given the arguments n and u, the RobustCAL$_\delta$ algorithm requests at most n labels, and if u is sufficiently large, and*

$$n \geq c_{k,t,a}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{k+1}\log\left(\frac{1}{\delta}\right),$$

then with probability at least $1-\delta$, the classifier $\hat{f} \in \mathcal{F}$ it returns satisfies $\mathrm{er}(\hat{f}) \leq \mathrm{er}(f^*) + \varepsilon$.

Corollary 26 proves an exponential label complexity speedup in the asymptotic dependence on ε compared to passive learning, for which there is a lower bound on the label complexity of $\Omega(1/\varepsilon)$ in the worst case over these distributions (Long, 1995).

Remark 27 *Condition 23 can be satisfied with $\alpha = 1$ if the Bayes optimal classifier is in $\mathcal{F}$ and the source distribution satisfies* Massart noise *(Massart and Nédélec, 2006):*

$$\Pr\left(|P(Y=1|X=x)-1/2| < 1/(2a)\right) = 0.$$

For example, if the data was generated by some unknown linear hypothesis with label noise (probability to flip any label) of up to $(a-1)/2a$, then P satisfies the requirements of Corollary 26.

Acknowledgements

R. El-Yaniv thanks the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI) and the Israel Science Foundation for their generous support.

Appendix A. Analysis of CAL via the Disagreement Coefficient

The following result was first established by (Giné and Koltchinskii, 2006, page 1213), with slightly different constant factors. The version stated here is directly from Hanneke (2009, Section 2.9), who also presents a simple and direct proof.

Lemma 28 (Giné and Koltchinskii, 2006; Hanneke, 2009) *For any $t \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{h \in \mathrm{VS}_{\mathcal{F},S_t}} \mathrm{er}(h) \leq \frac{24}{t}\left(d \ln\left(880 \cdot \theta(d/t)\right) + \ln\left(\frac{12}{\delta}\right)\right).$$

The following result is implicit in a proof of Hanneke (2011); for completeness, we present a formal proof here.

Lemma 29 (Hanneke, 2011) *There exists a finite universal constant $c_0 > 0$ such that, $\forall \delta \in (0,1)$, $\forall m \in \mathbb{N}$ with $m \geq 2$,*

$$\mathcal{B}_N(m,\delta) \leq c_0 \theta(d/m)\left(d \ln\left(e\theta(d/m)\right) + \ln\left(\frac{\log_2(m)}{\delta}\right)\right)\log_2(m).$$

Proof The result trivially holds for $m = 2$, taking any $c_0 \geq 2$. Otherwise, suppose $m \geq 3$. Note that, for any $t \in \mathbb{N}$,

$$\frac{24}{t}\left(d \ln\left(880\theta(d/t)\right) + \ln\left(\frac{24\log_2(m)}{\delta}\right)\right) \leq \frac{c_1}{t}\left(d \ln\left(e\theta(d/t)\right) + \ln\left(\frac{2\log_2(m)}{\delta}\right)\right), \quad (5)$$

for some universal constant $c_1 \in [1,\infty)$ (e.g., taking $c_1 = 168$ suffices). Thus, letting r_t denote the expression on the right hand side of (5), Lemma 28 implies that, for any $t \in \mathbb{N}$, with probability at least $1 - \delta/(2\log_2(m))$,

$$\sup_{h \in \mathrm{VS}_{\mathcal{F},S_t}} \mathrm{er}(h) \leq r_t.$$

By a union bound, this holds for all $t \in \{2^i : i \in \{1,\ldots,\lceil \log_2(m) \rceil - 1\}\}$ with probability at least $1-\delta/2$. In particular, on this event, we have

$$N(m;S_m) \leq 2 + \sum_{i=1}^{\lceil \log_2(m) \rceil - 1} \sum_{t=2^i+1}^{2^{i+1}} \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f^*,r_{2^i}))}(x_t).$$

A Chernoff bound implies that, with probability at least $1-\delta/2$, the right hand side is at most

$$\begin{aligned}
&\log_2\left(\frac{8}{\delta}\right)+2e\sum_{i=1}^{\lceil\log_2(m)\rceil-1}2^i\Delta\mathrm{B}(f^*,r_{2^i})\\
&\leq \log_2\left(\frac{8}{\delta}\right)+2e\sum_{i=1}^{\lceil\log_2(m)\rceil-1}2^i\theta\left(r_{2^i}\right)r_{2^i}\\
&\leq \log_2\left(\frac{8}{\delta}\right)+2ec_1\sum_{i=1}^{\lceil\log_2(m)\rceil-1}\theta\left(d2^{-i}\right)\left(d\ln\left(e\theta\left(d2^{-i}\right)\right)+\ln\left(\frac{2\log_2(m)}{\delta}\right)\right)\\
&\leq 4ec_1\theta(d/m)\left(d\ln\left(e\theta(d/m)\right)+\ln\left(\frac{\log_2(m)}{\delta}\right)\right)\log_2(m).
\end{aligned}$$

Letting $c_0=4ec_1$, the result holds by a union bound and minimality of $\mathcal{B}_N(m,\delta)$. ■

The following result is taken from the work of Hanneke (2011, Proof of Theorem 1); see also Hanneke (2014) for a theorem and proof expressed in this exact form.

Lemma 30 (Hanneke, 2011) *There exists a finite universal constant* $c_0>0$ *such that,* $\forall\varepsilon,\delta\in(0,1/2]$,

$$\Lambda(\varepsilon,\delta)\leq c_0\theta(\varepsilon)\left(d\ln(e\theta(\varepsilon))+\ln\left(\frac{\log_2(1/\varepsilon)}{\delta}\right)\right)\log_2\left(\frac{1}{\varepsilon}\right).$$

The next result is taken from the work of El-Yaniv and Wiener (2012, Corollary 39).

Lemma 31 (El-Yaniv and Wiener, 2012) *For any* $r_0\in(0,1)$,

$$\theta(r_0)\leq\max\left\{\sup_{r\in(r_0,1/2)}\frac{7\cdot\mathcal{B}_\Delta(\lfloor 1/r\rfloor,1/9)}{r},2\right\}.$$

Appendix B. Separation from the Previous Analyses

There are simple examples showing that sometimes $\mathcal{B}_{\hat{n}}(m,\delta)\approx\theta(1/m)$, so that the upper bound $\Lambda(\varepsilon,\delta)\leq c_0d\theta(\varepsilon)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$ in Lemma 30 is off by a factor of d compared to Theorem 10 in those cases (aside from logarithmic factors). For instance, consider the class of unions of k intervals, where $k\in\mathbb{N}$, $\mathcal{X}=[0,1]$, and $\mathcal{F}=\{x\mapsto 2\mathbb{1}_{\bigcup_{i=1}^k[z_{2i-1},z_{2i}]}(x)-1:0<z_1<\cdots<z_{2k}<1\}$. Suppose the data distribution P has a uniform marginal distribution over $\mathcal{X}$, and has $f^*=2\mathbb{1}_{\bigcup_{i=1}^k[z^*_{2i-1},z^*_{2i}]}-1$, where $z_i^*=\frac{i}{2k+1}$ for $i\in\{1,\ldots,2k\}$. In this case, for $r_0\geq 0$, $\theta(r_0)$ is within a factor of 2 of $\min\left\{\frac{1}{r_0},4k\right\}$ (see e.g., Balcan, Hanneke, and Vaughan, 2010; Hanneke, 2012). However, for any $m\in\mathbb{N}$ with $m\geq(2k+1)\ln\left(\frac{2k+1}{\delta}\right)$, with probability at least $1-\delta$ we have for each $i\in\{0,\ldots,2k\}$, at least one $j\leq m$ has $\frac{i}{2k+1}<x_j<\frac{i+1}{2k+1}$, and no $j\leq m$ has $x_j=\frac{i}{2k+1}$; in this case, $\hat{\mathcal{C}}_{S_m}$ is constructed as follows; for each $i\in\{1,\ldots,2k\}$, we include in $\hat{\mathcal{C}}_{S_m}$ the point (x_j,y_j) with largest x_j less than $\frac{i}{2k+1}$ and the point (x_j,y_j) with smallest x_j greater than $\frac{i}{2k+1}$. The number of points in this set $\hat{\mathcal{C}}_{S_m}$ is at most $4k$. Therefore, for any $m\in\mathbb{N}$, we have $\mathcal{B}_{\hat{n}}(m,\delta)\leq\min\left\{m,\max\left\{\left\lceil(2k+1)\ln\left(\frac{2k+1}{\delta}\right)\right\rceil,4k\right\}\right\}$. In particular, noting that $d=2k$ here, we have that for $\varepsilon<1/k$, the bound on $\Lambda(\varepsilon,\delta)$ in Lemma 30 has a $\tilde{\Theta}(k^2)$ dependence on k, while the upper bound on $\Lambda(\varepsilon,\delta)$ in Theorem 10 has only a $\tilde{\Theta}(k)$ dependence on k, which matches the lower bound in Theorem 10 (up to logarithmic factors).

Aside from the disagreement coefficient, the other technique in the existing literature for bounding the label complexity of CAL is due to El-Yaniv and Wiener (2010, 2012), based on a quantity they call the *characterizing set complexity*, denoted $\gamma(\mathcal{F},\hat{n}(S_m))$. Formally, for $n\in\mathbb{N}$, let $\gamma(\mathcal{F},n)$ denote the VC dimension of the collection of sets $\{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S}) : S\in(\mathcal{X}\times\mathcal{Y})^n\}$. Then El-Yaniv and Wiener (2012) prove the following bound, for a universal constant $c\in(0,\infty)$.[11]

$$\Lambda(\varepsilon,\delta)\le c\left(\max_{m\le M(\varepsilon,\delta/2)}\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))\ln\left(\frac{em}{\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))}\right)\right.$$
$$\left.+\ln\left(\frac{\log_2(2M(\varepsilon,\delta/2))}{\delta}\right)\right)\log_2(2M(\varepsilon,\delta/2)). \quad (6)$$

We can immediately note that $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))\ge\mathcal{B}_{\hat{n}}(m,\delta)-1$; specifically, for any $S\in(\mathcal{X}\times\mathcal{Y})^m$, letting $\{(x_{i_1},y_{i_1}),\ldots,(x_{i_{\hat{n}(S_m)}},y_{i_{\hat{n}(S_m)}})\}=\hat{\mathcal{C}}_S$, we have that $\{x_{i_2},\ldots,x_{i_{\hat{n}(S_m)}}\}$ is shattered by $\{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S'}) : S'\in(\mathcal{X}\times\mathcal{Y})^{\hat{n}(S_m)}\}$, since letting S' be any subset of $\{(x_{i_2},y_{i_2}),\ldots,(x_{i_{\hat{n}(S_m)}},y_{i_{\hat{n}(S_m)}})\}$ (filling in the remaining elements as copies of (x_{i_1},y_{i_1}) to make S' of size $\hat{n}(S_m)$),

$$\{(x_{i_2},y_{i_2}),\ldots,(x_{i_{\hat{n}(S_m)}},y_{i_{\hat{n}(S_m)}})\}\cap(\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S'})\times\mathcal{Y})=\{(x_{i_2},y_{i_2}),\ldots,(x_{i_{\hat{n}(S_m)}},y_{i_{\hat{n}(S_m)}})\}\setminus S',$$

since otherwise, the (x_{i_j},y_{i_j}) in $\{(x_{i_2},y_{i_2}),\ldots,(x_{i_{\hat{n}(S_m)}},y_{i_{\hat{n}(S_m)}})\}\setminus S'$ not in $\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S'})\times\mathcal{Y}$ would have $x_{i_j}\notin\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},\hat{\mathcal{C}}_S\setminus\{(x_{i_j},y_{i_j})\}})$, so that $\mathrm{VS}_{\mathcal{F},\hat{\mathcal{C}}_S\setminus\{(x_{i_j},y_{i_j})\}}=\mathrm{VS}_{\mathcal{F},\hat{\mathcal{C}}_S}=\mathrm{VS}_{\mathcal{F},S}$, contradicting minimality of $\hat{\mathcal{C}}_S$. Therefore, $\gamma(\mathcal{F},\hat{n}(S_m))\ge\hat{n}(S_m)-1$. Then noting that $\gamma(\mathcal{F},n)$ is monotonic in n, we find that $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))$ is a minimal $1-\delta$ confidence bound on $\gamma(\mathcal{F},\hat{n}(S_m))$, which implies $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))\ge\mathcal{B}_{\hat{n}}(m,\delta)-1$.

One can also give examples where the gap between $\mathcal{B}_{\hat{n}}(m,\delta)$ and $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta)$ is large, for instance where $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta))\ge d$ while $\mathcal{B}_{\hat{n}}(m,\delta)=2$ for large m. For instance, consider $\mathcal{X}$ that has d points $w_1,\ldots,w_d$ and 2^{d+1} additional points x_I and z_I indexed by the sets $I\subseteq\{1,\ldots,d\}$, and say $\mathcal{F}$ is the space of classifiers $\{h_J : J\subseteq\{1,\ldots,d\}\}$, where for each $J\subseteq\{1,\ldots,d\}$, $\{x : h_J(x)=+1\}=\{w_i : i\in J\}\cup\{x_I : I\subseteq J\}\cup\{z_I : I\subseteq\{1,\ldots,d\}\setminus J\}$; in particular, the classification on $w_1,\ldots,w_d$ determines the classification on the remaining 2^{d+1} points, and $\{w_1,\ldots,w_d\}$ is shatterable, so that $|\mathcal{F}|=2^d$, and the VC dimension of $\mathcal{F}$ is d. Let P be a distribution that has a uniform marginal distribution over the $2^{d+1}+d$ points in $\mathcal{X}$, and satisfies the realizable case assumption (i.e., $\mathbb{P}(Y=f^*(X)|X)=1$, for some $f^*\in\mathcal{F}$). For any integer $m\ge(2^{d+1}+d)\ln(2/\delta)$, with probability at least $1-\delta$, we have $(x_{\{i\le d:f^*(w_i)=+1\}},+1)\in S_m$ and $(z_{\{i\le d:f^*(w_i)=-1\}},+1)\in S_m$. Since every $h_J\in\mathcal{F}$ with $h_J(x_{\{i\le d:f^*(w_i)=+1\}})=+1$ has $\{i\le d : f^*(w_i)=+1\}\subseteq J=\{i\le d : h_J(w_i)=+1\}$, and every $h_J\in\mathcal{F}$ with $h_J(z_{\{i\le d:f^*(w_i)=-1\}})=+1$ has $\{i\le d : f^*(w_i)=-1\}\subseteq\{1,\ldots,d\}\setminus J=\{i\le d : h_J(w_i)=-1\}$, so that $\{i\le d : f^*(w_i)=+1\}\supseteq\{i\le d : h_J(w_i)=+1\}$, we have that every $h_J\in\mathcal{F}$ with both $h_J(x_{\{i\le d:f^*(w_i)=+1\}})=+1$ and $h_J(z_{\{i\le d:f^*(w_i)=-1\}})=+1$ has $\{i\le d : h_J(w_i)=+1\}=\{i\le d : f^*(w_i)=+1\}$. Since classifiers in $\mathcal{F}$ are completely determined by their classification of $\{w_1,\ldots,w_d\}$, this implies $h_J=f^*$. Therefore, letting $\hat{\mathcal{C}}_{S_m}=\{(x_{\{i\le d:f^*(w_i)=+1\}},+1),(z_{\{i\le d:f^*(w_i)=-1\}},+1)\}$, we have $\mathrm{VS}_{\mathcal{F},\hat{\mathcal{C}}_{S_m}}=\mathrm{VS}_{\mathcal{F},S_m}$, so that $\hat{n}(S_m)\le 2$ (in

11. This result can be derived from their Theorem 15 via reasoning analogous to the derivation of Theorem 10 from Lemma 8 above.

fact, one can easily show $\hat{n}(S_m) = 2$ in this case). Thus, for large m, $\mathcal{B}_{\hat{n}}(m,\delta) \leq 2$. However, for any $I \subseteq \{1,\ldots,d\}$, letting $S = \{(x_{\{1,\ldots,d\}\setminus I}, +1)\}$, we have $h_{\{1,\ldots,d\}\setminus I} \in \mathrm{VS}_{\mathcal{F},S}$, every $h \in \mathrm{VS}_{\mathcal{F},S}$ has $h(w_i) = +1$ for every $i \in \{1,\ldots,d\} \setminus I$, and every $i \in I$ has $h_{(\{1,\ldots,d\}\setminus I)\cup\{i\}} \in \mathrm{VS}_{\mathcal{F},S}$, so that $\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S}) \cap \{w_1,\ldots,w_d\} = \{w_i : i \in I\}$; therefore, the VC dimension of $\{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},\{x\}}) : x \in \mathcal{X}\}$ is at least d: that is, $\gamma(\mathcal{F},1) \geq d$. Since we have $\hat{n}(S_m) \geq 1$ whenever S_m contains any point other than $x_{\{\}}$ and $z_{\{\}}$, and this happens with probability at least $1-(2/(2^{d+1}+d))^m \geq 1-\delta > \delta$ (when $\delta < 1/2$), this implies we have $\gamma(\mathcal{F},\hat{n}(S_m)) \geq \gamma(\mathcal{F},1) \geq d$ with probability greater than δ, which (by monotonicity of $\gamma(\mathcal{F},\cdot)$) implies $\gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta)) \geq d$.

This is not quite strong enough to show a gap between (6) and Theorem 10, since the bounds in Theorem 10 require us to *maximize* over the value of m, which would therefore also include values $\mathcal{B}_{\hat{n}}(m,\delta)$ for $m < (2^{d+1}+d)\ln(2/\delta)$. To exhibit a gap between these bounds, we can simply redefine the marginal distribution of P over $\mathcal{X}$ to have $P(\{w_1\} \times \mathcal{Y}) = 1$. Note that with this distribution, $x_i = w_1$ for all i, with probability 1, so that we clearly have $\hat{n}(S_m) = 1$ almost surely, and hence $\mathcal{B}_{\hat{n}}(m,\delta) = 1$ for all m. As argued above, we have $\gamma(\mathcal{F},1) \geq d$ for this space. Therefore, $\max_{m \leq M} \gamma(\mathcal{F},\mathcal{B}_{\hat{n}}(m,\delta)) \geq d$, while $\max_{m \leq M} \mathcal{B}_{\hat{n}}(m,\delta) \leq 1$, for all $M \in \mathbb{N}$. However, note that unlike the example constructed above for the disagreement coefficient, the gap in this example could potentially be eliminated by replacing the distribution-free quantity $\gamma(\mathcal{F},n)$ with a distribution-dependent complexity measure (e.g., an annealed VC entropy or a bracketing number for $\{\mathrm{DIS}(\mathrm{VS}_{\mathcal{F},S}) : S \in (\mathcal{X} \times \mathcal{Y})^n\}$).

References

K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

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

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. In *Proceedings of the* 17th *Conference on Learning Theory*, 2004.

M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the* 26th *Conference on Learning Theory*, 2013.

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

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the* 20th *Conference on Learning Theory*, 2007.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the* 26th *International Conference on Machine Learning*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, 2010.

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

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

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. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13:255–279, 2012.

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

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

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

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

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

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of Active Learning. `http://www.stevehanneke.com`, 2014.

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012.

T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the* 8^{th} *Conference on Computational Learning Theory*, 1995.

L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *Journal of the Association for Computing Machinery*, 43(5):840–862, 1996.

R. Herbrich. *Learning Kernel Classifiers*. The MIT Press. Cambridge, MA, 2002.

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1975.

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

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

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27: 1808–1829, 1999.

P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006.

T. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the* 5$^{\text{th}}$ *International Joint Conference on Artificial Intelligence*, 1977.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.

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.

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011.

Y. Wiener. *Theoretical Foundations of Selective Prediction*. PhD thesis, the Technion — Israel Institute of Technology, 2013.

Y. Wiener and R. El-Yaniv. Pointwise tracking the optimal regression function. In *Advances in Neural Information Processing Systems 25*, 2012.

Y. Wiener and R. El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Machine Learning Research*, to appear, 2015.