Diameter-Based Active Learning

Christopher Tosh [1] **Sanjoy Dasgupta** [1]

Abstract

To date, the tightest upper and lower-bounds for the active learning of general concept classes have been in terms of a parameter of the learning problem called the *splitting index*. We provide, for the first time, an efficient algorithm that is able to realize this upper bound, and we empirically demonstrate its good performance.

1. Introduction

In many situations where a classifier is to be learned, it is easy to collect unlabeled data but costly to obtain labels. This has motivated the *pool-based active learning* model, in which a learner has access to a collection of unlabeled data points and is allowed to ask for individual labels in an adaptive manner. The hope is that choosing these queries intelligently will rapidly yield a low-error classifier, much more quickly than with random querying. A central focus of active learning is developing efficient querying strategies and understanding their label complexity.

Over the past decade or two, there has been substantial progress in developing such rigorously-justified active learning schemes for general concept classes. For the most part, these schemes can be described as *mellow*: rather than focusing upon maximally informative points, they query any point whose label cannot reasonably be inferred from the information received so far. It is of interest to develop more aggressive strategies with better label complexity.

An exception to this general trend is the aggressive strategy of (Dasgupta, 2005), whose label complexity is known to be optimal in its dependence on a key parameter called the *splitting index*. However, this strategy has been primarily of theoretical interest because it is difficult to implement algorithmically. In this paper, we introduce a variant of the methodology that yields efficient algorithms. We show that it admits roughly the same label complexity bounds as well as having promising experimental performance.

As with the original splitting index result, we operate in the *realizable* setting, where data can be perfectly classified by some function h^* in the hypothesis class $\mathcal{H}$. At any given time during the active learning process, the remaining candidates—that is, the elements of $\mathcal{H}$ consistent with the data so far—are called the *version space*. The goal of aggressive active learners is typically to pick queries that are likely to shrink this version space rapidly. But what is the right notion of size? Dasgupta (2005) pointed out that the *diameter* of the version space is what matters, where the distance between two classifiers is taken to be the fraction of points on which they make different predictions. Unfortunately, the diameter is a difficult measure to work with because it cannot, in general, be decreased at a steady rate. Thus the earlier work used a procedure that has quantifiable label complexity but is not conducive to implementation.

We take a fresh perspective on this earlier result. We start by suggesting an alternative, but closely related, notion of the size of a version space: the *average* pairwise distance between hypotheses in the version space, with respect to some underlying probability distribution π on $\mathcal{H}$. This distribution π can be arbitrary—that is, there is no requirement that the target h^* is chosen from it—but should be chosen so that it is easy to sample from. When $\mathcal{H}$ consists of linear separators, for instance, a good choice would be a log-concave density, such as a Gaussian.

At any given time, the next query x is chosen roughly as follows:

- Sample a collection of classifiers $h_1, h_2, \ldots, h_m$ from π restricted to the current version space V.
- Compute the distances between them; this can be done using just the unlabeled points.
- Any candidate query x partitions the classifiers $\{h_i\}$ into two groups: those that assign it a $+$ label (call these V_x^+) and those that assign it a $-$ label (call these V_x^-). Estimate the average-diameter after labeling x by the sum of the distances between classifiers h_i within V_x^+, or those within V_x^-, whichever is larger.
- Out of the pool of unlabeled data, pick the x for which

[1]Department of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA. Correspondence to: Christopher Tosh <ctosh@cs.ucsd.edu>, Sanjoy Dasgupta <dasgupta@cs.ucsd.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

this diameter-estimate is smallest.

This is repeated until the version space has small enough average diameter that a random sample from it is very likely to have error less than a user-specified threshold ϵ. We show how all these steps can be achieved efficiently, as long as there is a sampler for π.

Dasgupta (2005) pointed out that the label complexity of active learning depends on the underlying distribution, the amount of unlabeled data (since more data means greater potential for highly-informative points), and also the target classifier h^*. That paper identifies a parameter called the *splitting index* ρ that captures the relevant geometry, and gives upper bounds on label complexity that are proportional to $1/\rho$, as well as showing that this dependence is inevitable. For our modified notion of diameter, a different *averaged* splitting index is needed. However, we show that it can be bounded by the original splitting index, with an extra multiplicative factor of $\log(1/\epsilon)$; thus all previously-obtained label complexity results translate immediately for our new algorithm.

2. Related Work

The theory of active learning has developed along several fronts.

One of these is *nonparametric* active learning, where the learner starts with a pool of unlabeled points, adaptively queries a few of them, and then fills in the remaining labels. The goal is to do this with as few errors as possible. (In particular, the learner does not return a classifier from some predefined parametrized class.) One scheme begins by building a neighborhood graph on the unlabeled data, and propagating queried labels along the edges of this graph (Zhu et al., 2003; Cesa-Bianchi et al., 2009; Dasarathy et al., 2015). Another starts with a hierarchical clustering of the data and moves down the tree, sampling at random until it finds clusters that are relatively pure in their labels (Dasgupta & Hsu, 2008). The label complexity of such methods have typically be given in terms of smoothness properties of the underlying data distribution (Castro & Nowak, 2008; Kpotufe et al., 2015).

Another line of work has focused on active learning of linear separators, by querying points close to the current guess at the decision boundary (Balcan et al., 2007; Dasgupta et al., 2009; Balcan & Long, 2013). Such algorithms are close in spirit to those used in practice, but their analysis to date has required fairly strong assumptions to the effect that the underlying distribution on the unlabeled points is logconcave. Interestingly, regret guarantees for online algorithms of this sort can be shown under far weaker conditions (Cesa-Bianchi et al., 2006).

The third category of results, to which the present paper belongs, considers active learning strategies for general concept classes $\mathcal{H}$. Some of these schemes (Cohn et al., 1994; Dasgupta et al., 2007; Beygelzimer et al., 2009; Balcan et al., 2009; Zhang & Chaudhuri, 2014) are fairly mellow in the sense described earlier, using generalization bounds to gauge which labels can be inferred from those obtained so far. The label complexity of these methods can be bounded in terms of a quantity known as the disagreement coefficient (Hanneke, 2007). In the realizable case, the canonical such algorithm is that of (Cohn et al., 1994), henceforth referred to as CAL. Other methods use a prior distribution π over the hypothesis class, sometimes assuming that the target classifier is a random draw from this prior. These methods typically aim to shrink the mass of the version space under π, either greedily and explicitly (Dasgupta, 2004; Guillory & Bilmes, 2009; Golovin et al., 2010) or implicitly (Freund et al., 1997). Perhaps the most widely-used of these methods is the latter, query-by-committee, henceforth QBC. As mentioned earlier, shrinking π-mass is not an optimal strategy if low misclassification error is the ultimate goal. In particular, what matters is not the prior mass of the remaining version space, but rather how *different* these candidate classifiers are from each other. This motivates using the diameter of the version space as a yardstick, which was first proposed in (Dasgupta, 2005) and is taken up again here.

3. Preliminaries

Consider a binary hypothesis class $\mathcal{H}$, a data space $\mathcal{X}$, and a distribution $\mathcal{D}$ over $\mathcal{X}$. For mathematical convenience, we will restrict ourselves to finite hypothesis classes. (We can do this without loss of generality when $\mathcal{H}$ has finite VC dimension, since we only use the predictions of hypotheses on a pool of unlabeled points; however, we do not spell out the details of this reduction here.) The *hypothesis distance* induced by $\mathcal{D}$ over $\mathcal{H}$ is the pseudometric

$$d(h, h') := Pr_{x\sim\mathcal{D}}(h(x) \neq h'(x)).$$

Given a point $x \in \mathcal{X}$ and a subset $V \subset \mathcal{H}$, denote

$$V_x^+ = \{h \in V \,:\, h(x) = 1\}$$

and $V_x^- = V \setminus V_x^+$. Given a sequence of data points $x_1, \ldots, x_n$ and a target hypothesis h^*, the induced *version space* is the set of hypotheses that are consistent with the target hypotheses on the sequence, i.e.

$$\{h \in \mathcal{H} \,:\, h(x_i) = h^*(x_i) \text{ for all } i = 1, \ldots, n\}.$$

3.1. Diameter and the Splitting Index

The *diameter* of a set of hypotheses $V \subset \mathcal{H}$ is the maximal distance between any two hypotheses in V, i.e.

$$\text{diam}(V) := \max_{h,h'\in V} d(h, h').$$

Without any prior information, any hypothesis in the version space could be the target. Thus the worst case error of any hypothesis in the version space is the diameter of the version space. The splitting index roughly characterizes the number of queries required for an active learning algorithm to reduce the diameter of the version space below ϵ.

While reducing the diameter of a version space $V \subset \mathcal{H}$, we will sometimes identify pairs of hypotheses $h, h' \in V$ that are far apart and therefore need to be separated. We will refer to $\{h, h'\}$ as an *edge*. Given a set of edges $E = \{\{h_1, h_1'\}, \ldots, \{h_n, h_n'\}\} \subset \binom{\mathcal{H}}{2}$, we say a data point x ρ-splits E if querying x separates at least a ρ fraction of the pairs, that is, if

$$\max\left\{\left|E_x^+\right|, \left|E_x^-\right|\right\} \leq (1-\rho)|E|$$

where $E_x^+ = E \cap \binom{\mathcal{H}_x^+}{2}$ and similarly for E_x^-. When attempting to get accuracy $\epsilon > 0$, we need to only eliminate edge of length greater than ϵ. Define

$$E_\epsilon = \{\{h, h'\} \in E \,:\, d(h, h') > \epsilon\}.$$

The *splitting index* of a set $V \subset \mathcal{H}$ is a tuple (ρ, ϵ, τ) such that for all finite edge-sets $E \subset \binom{V}{2}$,

$$Pr_{x\sim\mathcal{D}}(x \,\rho\text{-splits } E_\epsilon) \geq \tau.$$

The following theorem, due to Dasgupta (2005), bounds the sample complexity of active learning in terms of the splitting index. The $\tilde{O}$ notation hides polylogarithmic factors in d, ρ, τ, $\log 1/\epsilon$, and the failure probability δ.

Theorem 1 (Dasgupta 2005)**.** *Suppose $\mathcal{H}$ is a hypothesis class with splitting index (ρ, ϵ, τ). Then to learn a hypothesis with error ϵ,*

(a) any active learning algorithm with $\leq 1/\tau$ unlabeled samples must request at least $1/\rho$ labels, and

(b) if $\mathcal{H}$ has VC-dimension d, there is an active learning algorithm that draws $\tilde{O}(d/(\rho\tau)\log^2(1/\epsilon))$ unlabeled data points and requests $\tilde{O}((d/\rho)\log^2(1/\epsilon))$ labels.

Unfortunately, the only known algorithm satisfying (b) above is intractable for all but the simplest hypothesis classes: it constructs an ϵ-covering of the hypothesis space and queries points which whittle away at the diameter of this covering. To overcome this intractability, we consider a slightly more benign setting in which we have a samplable prior distribution π over our hypothesis space $\mathcal{H}$.

3.2. An Average Notion of Diameter

With a prior distribution, it makes sense to shift away from the worst-case to the average-case. We define the *average diameter* of a subset $V \subset \mathcal{H}$ as the expected distance between two hypotheses in V randomly drawn from π, i.e.

$$\Phi(V) := \mathbb{E}_{h,h'\sim\pi|_V}[d(h, h')]$$

where $\pi|_V$ is the conditional distribution induced by restricting π to V, that is, $\pi|_V(h) = \pi(h)/\pi(V)$ for $h \in V$.

Intuitively, a version space with very small average diameter ought to put high weight on hypotheses that are close to the true hypothesis. Indeed, given a version space V with $h^* \in V$, the following lemma shows that if $\Phi(V)$ is small enough, then a low error hypothesis can be found by two popular heuristics: random sampling and MAP estimation.

Lemma 2. *Suppose $V \subset \mathcal{H}$ contains h^*. Pick $\epsilon > 0$.*

(a) (Random sampling) If $\Phi(V) \leq \epsilon\,\pi|_V(h^*)$ then $\mathbb{E}_{h\sim\pi|_V}[d(h^*, h)] \leq \epsilon$.

(b) (MAP estimation) Write $p_{map} = \max_{h\in V}\pi|_V(h)$. Pick $0 < \alpha < p_{map}$. If

$$\Phi(V) \leq 2\epsilon\left(\min\{\pi|_V(h^*), p_{map} - \alpha\}\right)^2,$$

then $d(h^*, h) \leq \epsilon$ for any h with $\pi|_V(h) \geq p_{map} - \alpha$.

Proof. Part (a) follows from

$$\Phi(V) = \mathbb{E}_{h,h'\sim\pi|_V}[d(h, h')] \geq \pi|_V(h^*)\mathbb{E}_{h\sim\pi|_V}[d(h^*, h)].$$

For (b), take $\delta = \min(\pi|_V(h^*), p_{map} - \alpha)$ and define $V_{\pi,\delta} = \{h \in V \,:\, \pi|_V(h) \geq \delta\}$. Note that $V_{\pi,\delta}$ contains h^* as well as any $h \in V$ with $\pi|_V(h) \geq p_{map} - \alpha$.

We claim $\mathrm{diam}(V_{\pi,\delta})$ is at most ϵ. Suppose not. Then there exist $h_1, h_2 \in V_{\pi,\delta}$ satisfying $d(h_1, h_2) > \epsilon$, implying

$$\begin{aligned}\Phi(V) &= \mathbb{E}_{h,h'\sim\pi|_V}[d(h, h')] \\ &\geq 2 \cdot \pi|_V(h_1) \cdot \pi|_V(h_2) \cdot d(h_1, h_2) > 2\delta^2\epsilon.\end{aligned}$$

But this contradicts our assumption on $\Phi(V)$. Since both $h, h^* \in V_{\pi,\delta}$, we have (b). □

3.3. An Average Notion of Splitting

We now turn to defining an average notion of splitting. A data point x *ρ-average splits* V if

$$\max\left\{\frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+), \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-)\right\} \leq (1-\rho)\Phi(V).$$

And we say a set $S \subset \mathcal{H}$ has *average splitting index* (ρ, ϵ, τ) if for any subset $V \subset S$ such that $\Phi(V) > \epsilon$,

$$Pr_{x\sim\mathcal{D}}\left(x \,\rho\text{-average splits } V\right) \geq \tau.$$

Intuitively, average splitting refers to the ability to significantly decrease the potential function

$$\pi(V)^2\Phi(V) = \mathbb{E}_{h,h'\sim\pi}[\mathbb{1}(h,h' \in V)\, d(h,h')]$$

with a single query.

While this potential function may seem strange at first glance, it is closely related to the original splitting index. The following lemma, whose proof is deferred to Section 5, shows the splitting index bounds the average splitting index for any hypothesis class.

Lemma 3. *Let π be a probability measure over a hypothesis class $\mathcal{H}$. If $\mathcal{H}$ has splitting index (ρ, ϵ, τ), then it has average splitting index $(\frac{\rho}{4\lceil\log(1/\epsilon)\rceil}, 2\epsilon, \tau)$.*

Dasgupta (2005) derived the splitting indices for several hypothesis classes, including intervals and homogeneous linear separators. Lemma 3 implies average splitting indices within a $\log(1/\epsilon)$ factor in these settings.

Moreover, given access to samples from $\pi|_V$, we can easily estimate the quantities appearing in the definition of average splitting. For an edge sequence $E = (\{h_1, h_1'\}, \ldots, \{h_n, h_n'\})$, define

$$\psi(E) := \sum_{i=1}^{n} d(h_i, h_i').$$

When h_i, h_i' are i.i.d. draws from $\pi|_V$ for all $i = 1, \ldots, n$, which we denote $E \sim (\pi|_V)^{2\times n}$, the random variables $\psi(E)$, $\psi(E_x^-)$, and $\psi(E_x^+)$ are unbiased estimators of the quantities appearing in the definition of average splitting.

Lemma 4. *Given $E \sim (\pi|_V)^{2\times n}$, we have*

$$\mathbb{E}\left[\frac{1}{n}\psi(E)\right] = \Phi(V) \text{ and } \mathbb{E}\left[\frac{1}{n}\psi(E_x^+)\right] = \frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+)$$

for any $x \in \mathcal{X}$. Similarly for E_x^- and V_x^-.

Proof. From definitions and linearity of expectations, it is easy to observe $\mathbb{E}[\psi(E)] = n\,\Phi(V)$. By the independence of h_i, h_i', we additionally have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\psi(E_x^+)\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{\{h_i,h_i'\}\in E_x^+} d(h_i,h_i')\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{\{h_i,h_i'\}\in E} \mathbb{1}[h_i \in V_x^+]\,\mathbb{1}[h_i' \in V_x^+]\, d(h_i,h_i')\right] \\
&= \frac{1}{n}\sum_{\{h_i,h_i'\}\in E}\left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \mathbb{E}\left[d(h_i,h_i') \,|\, h_i, h_i' \in V_x^+\right] \\
&= \left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \Phi(V_x^+). \qquad \square
\end{aligned}$$

Remark: It is tempting to define average splitting in terms of the average diameter as

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \le (1-\rho)\Phi(V).$$

However, this definition does not satisfy a nice relationship with the splitting index. Indeed, there exist hypothesis classes V for which there are many points which $1/4$-split E for any $E \subset \binom{V}{2}$ but for which every $x \in \mathcal{X}$ satisfies

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \approx \Phi(V).$$

This observation is formally proven in the appendix.

4. An Average Splitting Index Algorithm

Suppose we are given a version space V with average splitting index (ρ, ϵ, τ). If we draw $\tilde{O}(1/\tau)$ points from the data distribution then, with high probability, one of these will ρ-average split V. Querying that point will result in a version space V' with significantly smaller potential $\pi(V')^2\Phi(V')$.

If we knew the value ρ a priori, then Lemma 4 combined with standard concentration bounds (Hoeffding, 1963; Angluin & Valiant, 1977) would give us a relatively straightforward procedure to find a good query point:

1. Draw $E' \sim (\pi|_V)^{2\times M}$ and compute the empirical estimate $\widehat{\Phi}(V) = \frac{1}{M}\psi(E')$.
2. Draw $E \sim (\pi|_V)^{2\times N}$ for N depending on ρ and $\widehat{\Phi}$.
3. For suitable M and N, it will be the case that with high probability, for some x,

$$\frac{1}{N}\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} \approx (1-\rho)\widehat{\Phi}.$$

   Querying that point will decrease the potential.

However, we typically would not know the average splitting index ahead of time. Moreover, it is possible that the average splitting index may change from one version space to the next. In the next section, we describe a query selection procedure that adapts to the splittability of the current version space.

4.1. Finding a Good Query Point

Algorithm 2, which we term SELECT, is our query selection procedure. It takes as input a sequence of data points $x_1, \ldots, x_m$, at least one of which ρ-average splits the current version space, and with high probability finds a data point that $\rho/8$-average splits the version space.

SELECT proceeds by positing an optimistic estimate of ρ, which we denote $\widehat{\rho}_t$, and successively halving it until we are

Algorithm 1 DBAL

Input: Hypothesis class $\mathcal{H}$, prior distribution π
Initialize $V = \mathcal{H}$
while $\frac{1}{n}\psi(E) \geq \frac{3\epsilon}{4}$ for $E \sim (\pi|_V)^{2\times n}$ **do**
  Draw m data points $\mathbf{x} = (x_1, \ldots, x_m)$
  Query point $x_i = \text{SELECT}(V, \mathbf{x})$ and set V to be consistent with the result
end while
return Current version space V in the form of the queried points $(x_1, h^*(x_1)), \ldots, (x_K, h^*(x_K))$

Algorithm 2 SELECT

Input: Version space V, prior π, data $\mathbf{x} = (x_1, \ldots, x_m)$
Set $\widehat{\rho}_1 = 1/2$
for $t = 1, 2, \ldots$ **do**
  Draw $E' \sim (\pi|_V)^{2\times m_t}$ and compute $\widehat{\Phi}_t = \frac{1}{m_t}\psi(E')$
  Draw $E \sim (\pi|_V)^{2\times n_t}$
  If $\exists\, x_i$ s.t. $\frac{1}{n_t}\max\left\{\psi(E^+_{x_i}), \psi(E^-_{x_i})\right\} \leq (1-\widehat{\rho}_t)\widehat{\Phi}_t$, then **halt** and **return** x_i
  Otherwise, let $\widehat{\rho}_{t+1} = \widehat{\rho}_t/2$
end for

confident that we have found a point that $\widehat{\rho}_t$-average splits the version space. In order for this algorithm to succeed, we need to choose n_t and m_t such that with high probability (1) $\widehat{\Phi}_t$ is an accurate estimate of $\Phi(V)$ and (2) our halting condition will be true if $\widehat{\rho}_t$ is within a constant factor of ρ and false otherwise. The following lemma, whose proof is in the appendix, provides such choices for n_t and m_t.

Lemma 5. *Let $\rho, \epsilon, \delta_0 > 0$ be given. Suppose that version space V satisfies $\Phi(V) > \epsilon$. In* SELECT*, fix a round t and data point $x \in \mathcal{X}$ that exactly ρ-average splits V (that is, $\max\{\pi|_V(V_x^+)^2\Phi(V_x^+),\ \pi|_V(V_x^-)^2\Phi(V_x^-)\} = (1-\rho)\Phi(V)$). If*

$$m_t \geq \frac{48}{\widehat{\rho}_t^2\epsilon}\log\frac{4}{\delta_0} \text{ and } n_t \geq \max\left\{\frac{32}{\widehat{\rho}_t^2\widehat{\Phi}_t}, \frac{40}{\widehat{\Phi}_t^2}\right\}\log\frac{4}{\delta_0}$$

then with probability $1-\delta_0$, $\widehat{\Phi}_t \geq (1-\widehat{\rho}_t/4)\Phi(V)$ and

(a) if $\rho \leq \widehat{\rho}_t/2$, then

$$\frac{1}{n_t}\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} > (1-\widehat{\rho}_t)\widehat{\Phi}_t.$$

(b) If $\rho \geq 2\widehat{\rho}_t$, then

$$\frac{1}{n_t}\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} \leq (1-\widehat{\rho}_t)\widehat{\Phi}_t.$$

Given the above lemma, we can establish a bound on the number of rounds and the total number of hypotheses SELECT needs to find a data point that $\rho/8$-average splits the version space.

Theorem 6. *Suppose that* SELECT *is called with a version space V with $\Phi(V) \geq \epsilon$ and a collection of points $x_1, \ldots, x_m$ such that at least one of x_i ρ-average splits V. If $\delta_0 \leq \delta/(2m(2+\log(1/\rho)))$, then with probability at least $1-\delta$,* SELECT *returns a point x_i that $(\rho/8)$-average splits V, finishing in less than $\lceil\log(1/\rho)\rceil + 1$ rounds and sampling $O\left(\left(\frac{1}{\epsilon\rho^2} + \frac{\log(1/\rho)}{\Phi(V)^2}\right)\log\frac{1}{\delta_0}\right)$ hypotheses in total.*

Remark 1: It is possible to modify SELECT to find a point x_i that $(c\rho)$-average splits V for any constant $c < 1$ while only having to draw $O(1)$ more hypotheses in total. First note that by halving $\widehat{\rho}_t$ at each step, we immediately give up a factor of two in our approximation. This can be made smaller by taking narrower steps. Additionally, with a constant factor increase in m_t and n_t, the approximation ratios in Lemma 5 can be set to any constant.

Remark 2: At first glance, it appears that SELECT requires us to know ρ in order to calculate δ_0. However, a crude lower bound on ρ suffices. Such a bound can always be found in terms of ϵ. This is because any version space is $(\epsilon/2, \epsilon, \epsilon/2)$-splittable (Dasgupta, 2005, Lemma 1). By Lemma 3, so long as τ is less than $\epsilon/4$, we can substitute $\frac{\epsilon}{8\lceil\log(2/\epsilon)\rceil}$ for ρ in when we compute δ_0.

Proof of Theorem 6. Let $T := \lceil\log(1/\rho)\rceil + 1$. By Lemma 5, we know that for rounds $t = 1, \ldots, T$, we don't return any point which does worse than $\widehat{\rho}_t/2$-average splits V with probability $1-\delta/2$. Moreover, in the T-th round, it will be the case that $\rho/4 \leq \widehat{\rho}_T \leq \rho/2$, and therefore, with probability $1-\delta/2$, we will select a point which does no worse than $\widehat{\rho}_T/2$-average split V, which in turn does no worse than $\rho/8$-average split V.

Note that we draw $m_t + n_t$ hypotheses at each round. By Lemma 5, for each round $\widehat{\Phi}_t \geq 3\Phi(V)/4 \geq 3\epsilon/4$. Thus

$$\begin{aligned}
\#\text{ of hypotheses drawn } &= \sum_{t=1}^{T} m_t + n_t \\
&= \sum_{t=1}^{T}\left(\frac{48}{\widehat{\rho}_t^2\epsilon} + \frac{32}{\widehat{\rho}_t^2\widehat{\Phi}_t} + \frac{40}{\widehat{\Phi}_t^2}\right)\log\frac{4}{\delta_0} \\
&\leq \sum_{t=1}^{T}\left(\frac{96}{\epsilon\widehat{\rho}_t^2} + \frac{72}{\Phi(V)^2}\right)\log\frac{4}{\delta_0}
\end{aligned}$$

Given $\widehat{\rho}_t = 1/2^t$ and $T \leq 2 + \log 1/\rho$, we have

$$\sum_{t=1}^{T}\frac{1}{\widehat{\rho}_t^2} = \sum_{t=1}^{T}2^{2t} \leq \left(\sum_{t=1}^{T}2^t\right)^2 \leq \left(2^{2+\log 1/\rho}\right)^2 = \frac{16}{\rho^2}.$$

Plugging in $\delta_0 \leq \frac{\delta}{2m(2+\log(1/\rho))}$, we recover the theorem statement. □

4.2. Active Learning Strategy

Using the SELECT procedure as a subroutine, Algorithm 1, henceforth DBAL for Diameter-based Active Learning, is our active learning strategy. Given a hypothesis class with average splitting index $(\rho, \epsilon/2, \tau)$, DBAL queries data points provided by SELECT until it is confident $\Phi(V) < \epsilon$.

Denote by V_t the version space in the t-th round of DBAL. The following lemma, which is proven in the appendix, demonstrates that the halting condition (that is, $\psi(E) < 3\epsilon n/4$, where E consists of n pairs sampled from $(\pi|_V)^2$) guarantees that with high probability DBAL stops when $\Phi(V_t)$ is small.

Lemma 7. *The following holds for DBAL:*

(a) Suppose that for all $t = 1, 2, \ldots, K$ that $\Phi(V_t) > \epsilon$. Then the probability that the termination condition is ever true for any of those rounds is bounded above by $K \exp\left(-\frac{\epsilon n}{32}\right)$.

(b) Suppose that for some $t = 1, 2, \ldots, K$ that $\Phi(V_t) \leq \epsilon/2$. Then the probability that the termination condition is not true in that round is bounded above by $K \exp\left(-\frac{\epsilon n}{48}\right)$.

Given the guarantees on the SELECT procedure in Theorem 6 and on the termination condition provided by Lemma 7, we get the following theorem.

Theorem 8. *Suppose that $\mathcal{H}$ has average splitting index $(\rho, \epsilon/2, \tau)$. Then DBAL returns a version space V satisfying $\Phi(V) \leq \epsilon$ with probability at least $1 - \delta$ while using the following resources:*

(a) $K \leq \frac{8}{\rho}\left(\log\frac{2}{\epsilon} + 2\log\frac{1}{\pi(h^*)}\right)$ rounds, with one label per round,

(b) $m \leq \frac{1}{\tau}\log\frac{2K}{\delta}$ unlabeled data points sampled per round, and

(c) $n \leq O\left(\left(\frac{1}{\epsilon\rho^2} + \frac{\log(1/\rho)}{\epsilon^2}\right)\left(\log\frac{mK}{\delta} + \log\log\frac{1}{\epsilon}\right)\right)$ hypotheses sampled per round.

Proof. From definition of the average splitting index, if we draw $m = \frac{1}{\tau}\log\frac{2K}{\delta}$ unlabeled points per round, then with probability $1 - \delta/2$, each of the first K rounds will have at least one data point that ρ-average splits the current version space. In each such round, if the version space has average diameter at least $\epsilon/2$, then with probability $1 - \delta/4$ SELECT will return a data point that $\rho/8$-average splits the current version space while sampling no more than $n = O\left(\left(\frac{1}{\epsilon\rho^2} + \frac{1}{\epsilon^2}\log\frac{1}{\rho}\right)\log\frac{mK\log\frac{1}{\epsilon}}{\delta}\right)$ hypotheses per round by Theorem 6.

By Lemma 7, if the termination check uses $n' = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ hypotheses per round, then with probability $1 - \delta/4$ in the first K rounds the termination condition will never be true when the current version space has average diameter greater than ϵ and will certainly be true if the current version space has diameter less than $\epsilon/2$.

Thus it suffices to bound the number of rounds in which we can $\rho/8$-average split the version space before encountering a version space with $\epsilon/2$.

Since the version space is always consistent with the true hypothesis h^*, we will always have $\pi(V_t) \geq \pi(h^*)$. After $K = \frac{8}{\rho}\left(\log\frac{2}{\epsilon} + 2\log\frac{1}{\pi(h^*)}\right)$ rounds of $\rho/8$-average splitting, we have

$$\begin{aligned}\pi(h^*)^2\Phi(V_K) &\leq \pi(V_K)^2\Phi(V_K)\\ &\leq \left(1 - \frac{\rho}{8}\right)^K \pi(V_0)^2\Phi(V_0)\\ &\leq \frac{\pi(h^*)^2\epsilon}{2}\end{aligned}$$

Thus in the first K rounds, we must terminate with a version space with average diameter less than ϵ. □

5. Proof of Lemma 3

In this section, we give the proof of the following relationship between the original splitting index and our average splitting index.

Lemma 3. *Let π be a probability measure over a hypothesis class $\mathcal{H}$. If $\mathcal{H}$ has splitting index (ρ, ϵ, τ), then it has average splitting index $(\frac{\rho}{4\lceil\log(1/\epsilon)\rceil}, 2\epsilon, \tau)$.*

The first step in proving Lemma 3 is to relate the splitting index to our estimator $\psi(\cdot)$. Intuitively, splittability says that for any set of large edges there are many data points which remove a significant fraction of them. One may suspect this should imply that if a set of edges is large on average, then there should be many data points which remove a significant fraction of their weight. The following lemma confirms this suspicion.

Lemma 9. *Suppose that $V \subset \mathcal{H}$ has splitting index (ρ, ϵ, τ), and say $E = (\{h_1, h_1'\}, \ldots, \{h_n, h_n'\})$ is a sequence of hypothesis pairs from V satisfying $\frac{1}{n}\psi(E) > 2\epsilon$. Then if $x \sim \mathcal{D}$, we have with probability at least τ,*

$$\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} \leq \left(1 - \frac{\rho}{4\lceil\log(1/\epsilon)\rceil}\right)\psi(E).$$

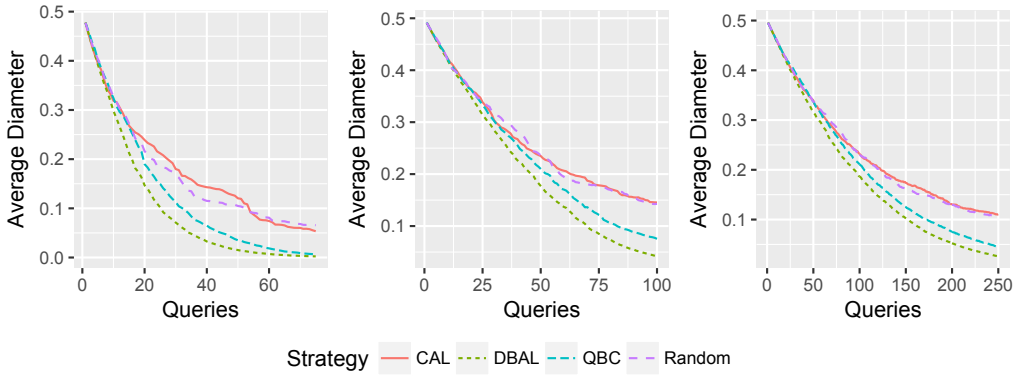

Figure 1. Simulation results on homogeneous linear separators. *Left*: $d = 10$. *Middle*: $d = 25$. *Right*: $d = 50$.

Proof. Consider partitioning E as

$$E_0 = \{\{h, h'\} \in E : d(h, h') < \epsilon\} \text{ and}$$
$$E_k = \{\{h, h'\} \in E : d(h, h') \in [2^{k-1}\epsilon, 2^k\epsilon)$$

for $k = 1, \ldots, K$ with $K = \lceil \log \frac{1}{\epsilon} \rceil$. Then $E_0, \ldots, E_K$ are all disjoint and their union is E. Define $E_{1:K} = \cup_{k=1}^K E_k$.

We first claim that $\psi(E_{1:K}) > \psi(E_0)$. This follows from the observation that because $\psi(E) \geq 2n\epsilon$ and each edge in E_0 has length less than ϵ, we must have

$$\psi(E_{1:K}) = \psi(E) - \psi(E_0) > 2n\epsilon - n\epsilon > \psi(E_0).$$

Next, observe that because each edge $\{h, h'\} \in E_k$ with $k \geq 1$ satisfies $d(h, h') \in [2^{k-1}\epsilon, 2^k\epsilon)$, we have

$$\psi(E_{1:K}) = \sum_{k=1}^{K} \sum_{\{h,h'\} \in E_k} d(h, h') \leq \sum_{k=1}^{K} 2^k \epsilon |E_k|.$$

Since there are only K summands on the right, at least one of these must be larger than $\psi(E_{1:K})/K$. Let k denote that index and let x be a point which ρ-splits E_k. Then we have

$$\begin{aligned}\psi((E_{1:K})_x^+) &\leq \psi(E_{1:K}) - \psi(E_k \setminus (E_k)_x^+) \\ &\leq \psi(E_{1:K}) - \rho 2^{k-1}\epsilon |E_k| \\ &\leq \left(1 - \frac{\rho}{2K}\right) \psi(E_{1:K}).\end{aligned}$$

Since $\psi(E_{1:K}) \geq \psi(E_0)$, we have

$$\begin{aligned}\psi(E_x^+) &\leq \psi(E_0) + \left(1 - \frac{\rho}{2K}\right) \psi(E_{1:K}) \\ &\leq \left(1 - \frac{\rho}{4K}\right) \psi(E).\end{aligned}$$

Symmetric arguments show the same holds for E_x^-.

Finally, by the definition of splitting, the probability of drawing a point x which ρ-splits E_k is at least τ, giving us the lemma. □

With Lemma 9 in hand, we are now ready to prove Lemma 3.

Proof of Lemma 3. Let $V \subset \mathcal{H}$ such that $\Phi(V) > 2\epsilon$. Suppose that we draw n edges E i.i.d. from $\pi|_V$ and draw a data point $x \sim \mathcal{D}$. Then Hoeffding's inequality (Hoeffding, 1963), combined with Lemma 4, tells us that there exist sequences $\epsilon_n, \delta_n \searrow 0$ such that with probability at least $1 - 3\delta_n$, the following hold simultaneously:

- $\Phi(V) - \epsilon_n \leq \frac{1}{n}\psi(E) \leq \Phi(V) + \epsilon_n$,
- $\frac{1}{n}\psi(E_x^+) \geq \frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+) - \epsilon_n$, and
- $\frac{1}{n}\psi(E_x^-) \geq \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-) - \epsilon_n$.

For ϵ_n small enough, we have that $\Phi(V) - \epsilon_n > 2\epsilon$. Combining the above with Lemma 9, we have with probability at least $\tau - 3\delta_n$,

$$\begin{aligned}&\max\left\{\frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+), \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-)\right\} - \epsilon_n \\ &\qquad \leq \frac{1}{n}\max\{\psi(E_x^+), \psi(E_x^-)\} \\ &\qquad \leq \left(1 - \frac{\rho}{4\lceil \log(1/\epsilon)\rceil}\right)\frac{\psi(E)}{n} \\ &\qquad \leq \left(1 - \frac{\rho}{4\lceil \log(1/\epsilon)\rceil}\right)(\Phi(V) + \epsilon_n)\end{aligned}$$

By taking $n \to \infty$, we have $\epsilon_n, \delta_n \searrow 0$, giving us the lemma. □

6. Simulations

We compared DBAL against the baseline passive learner as well as two other generic active learning strategies: CAL

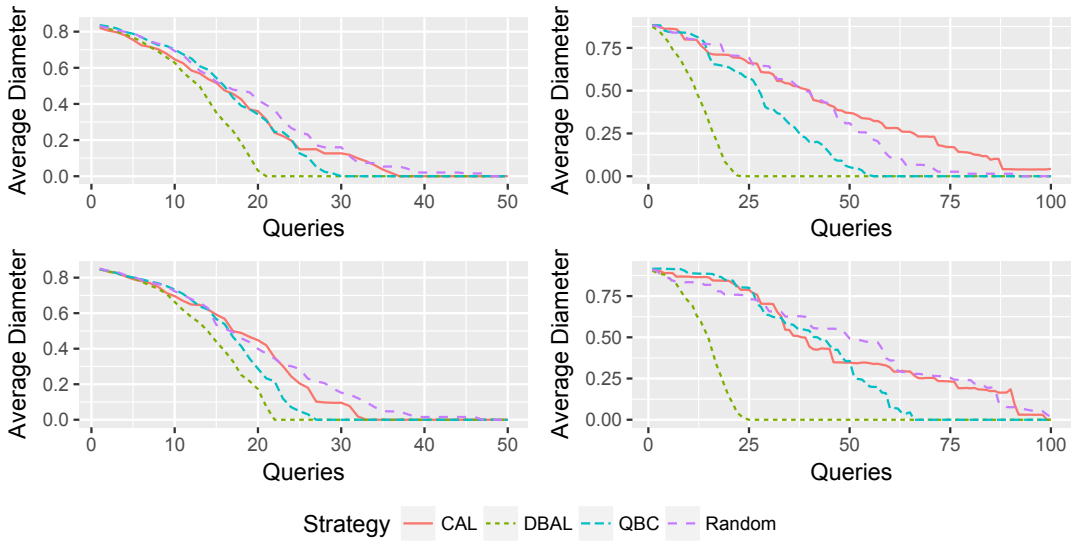

Figure 2. Simulation results on k-sparse monotone disjunctions. In all cases $k = 4$. *Top left*: $d = 75$, $p = 0.25$. *Top right*: $d = 75$, $p = 0.5$. *Bottom left*: $d = 100$, $p = 0.25$. *Bottom right*: $d = 100$, $p = 0.5$.

and QBC. CAL proceeds by randomly sampling a data point and querying it if its label cannot be inferred from previously queried data points. QBC uses a prior distribution π and maintains a version space V. Given a randomly sampled data point x, QBC samples two hypotheses $h, h' \sim \pi|_V$ and queries x if $h(x) \neq h'(x)$.

We tested on two hypothesis classes: homogeneous, or through-the-origin, linear separators and k-sparse monotone disjunctions. In each of our simulations, we drew our target h^* from the prior distribution. After each query, we estimated the average diameter of the version space. We repeated each simulation several times and plotted the average performance of each algorithm.

Homogeneous linear separators The class of d-dimensional homogeneous linear separators can be identified with elements of the d-dimensional unit sphere. That is, a hypothesis $h \in \mathcal{S}^{d-1}$ acts on a data point $x \in \mathbb{R}^d$ via the sign of their inner product:

$$h(x) := \text{sign}(\langle h, x \rangle).$$

In our simulations, both the prior distribution and the data distribution are uniform over the unit sphere. Although there is no known method to exactly sample uniformly from the version space, Gilad-Bachrach et al. (2005) demonstrated that using samples generated by the hit-and-run Markov chain works well in practice. We adopted this approach for our sampling tasks.

Figure 1 shows the results of our simulations on homogeneous linear separators.

Sparse monotone disjunctions A k-sparse monotone disjunction is a disjunction of k positive literals. Given a Boolean vector $x \in \{0, 1\}^n$, a monotone disjunction h classifies x as positive if and only if $x_i = 1$ for some positive literal i in h.

In our simulations, each data point is a vector whose coordinates are i.i.d. Bernoulli random variables with parameter p. The prior distribution is uniform over all k-sparse monotone disjunctions. When k is constant, it is possible to sample from the prior restricted to the version space in expected polynomial time using rejection sampling.

The results of our simulations on k-sparse monotone disjunctions are in Figure 2.

Acknowledgments

The authors are grateful to the reviewers for their feedback and to the NSF for support under grants IIS-1162581 and DGE-1144086. Part of this work was done at the Simons Institute for Theoretical Computer Science, Berkeley, as part of a program on the foundations of machine learning. CT additionally thanks Daniel Hsu and Stefanos Poulis for helpful discussions.

References

Angluin, Dana and Valiant, Leslie G. Fast probabilistic algorithms for hamiltonian circuits and matchings. In *Proceedings of the ninth annual ACM symposium on Theory of computing*, pp. 30–41. ACM, 1977.

Balcan, Maria-Florina and Long, Phil. Active and passive

learning of linear separators under log-concave distributions. In *Proceedings of the 26th Conference on Learning Theory*, pp. 288–316, 2013.

Balcan, Maria-Florina, Broder, Andrei, and Zhang, Tong. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Balcan, Maria-Florina, Beygelzimer, Alina, and Langford, John. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Beygelzimer, Alina, Dasgupta, Sanjoy, and Langford, John. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 49–56, 2009.

Castro, Rui M and Nowak, Robert D. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Cesa-Bianchi, Nicolo, Gentile, Claudio, and Zaniboni, Luca. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7:1205–1230, 2006.

Cesa-Bianchi, Nicolo, Gentile, Claudio, and Vitale, Fabio. Learning unknown graphs. In *International Conference on Algorithmic Learning Theory*, pp. 110–125. Springer, 2009.

Cohn, David, Atlas, Les, and Ladner, Richard. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

Dasarathy, Gautam, Nowak, Robert, and Zhu, Xiaojin. S2: An efficient graph based active learning algorithm with application to nonparametric classification. In *Proceedings of The 28th Conference on Learning Theory*, pp. 503–522, 2015.

Dasgupta, Sanjoy. Analysis of a greedy active learning strategy. In *Advances in neural information processing systems*, pp. 337–344, 2004.

Dasgupta, Sanjoy. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*, pp. 235–242, 2005.

Dasgupta, Sanjoy and Hsu, Daniel. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 208–215. ACM, 2008.

Dasgupta, Sanjoy, Monteleoni, Claire, and Hsu, Daniel J. A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pp. 353–360, 2007.

Dasgupta, Sanjoy, Kalai, Adam Tauman, and Monteleoni, Claire. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10(Feb):281–299, 2009.

Freund, Yoav, Seung, H Sebastian, Shamir, Eli, and Tishby, Naftali. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

Gilad-Bachrach, Ran, Navot, Amir, and Tishby, Naftali. Query by committee made real. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pp. 443–450. MIT Press, 2005.

Golovin, Daniel, Krause, Andreas, and Ray, Debajyoti. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pp. 766–774, 2010.

Guillory, Andrew and Bilmes, Jeff. Average-case active learning with costs. In *International Conference on Algorithmic Learning Theory*, pp. 141–155. Springer, 2009.

Hanneke, Steve. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 353–360. ACM, 2007.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Kpotufe, Samory, Urner, Ruth, and Ben-David, Shai. Hierarchical label queries with data-dependent partitions. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1176–1189, 2015.

Zhang, Chicheng and Chaudhuri, Kamalika. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pp. 442–450, 2014.

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.