Improved Algorithms for Agnostic Pool-based Active Classification

Julian Katz-Samuels [1] **Jifan Zhang** [2] **Lalit Jain** [2] **Kevin Jamieson** [2]

Abstract

We consider active learning for binary classification in the agnostic pool-based setting. The vast majority of works in active learning in the agnostic setting are inspired by the CAL algorithm where each query is uniformly sampled from the disagreement region of the current version space. The sample complexity of such algorithms is described by a quantity known as the disagreement coefficient which captures both the geometry of the hypothesis space as well as the underlying probability space. To date, the disagreement coefficient has been justified by minimax lower bounds only, leaving the door open for superior instance dependent sample complexities. In this work we propose an algorithm that, in contrast to uniform sampling over the disagreement region, solves an experimental design problem to determine a distribution over examples from which to request labels. We show that the new approach achieves sample complexity bounds that are never worse than the best disagreement coefficient-based bounds, but in specific cases can be dramatically smaller. From a practical perspective, the proposed algorithm requires no hyperparameters to tune (e.g., to control the aggressiveness of sampling), and is computationally efficient by means of assuming access to an empirical risk minimization oracle (without any constraints). Empirically, we demonstrate that our algorithm is superior to state of the art agnostic active learning algorithms on image classification datasets.

1. Introduction

Most applications of machine learning have an enormous amount of unlabeled data. Yet, many powerful machine learning methods require that this data be labeled and reliable labels are costly since they require human intervention. The cost of providing labels has become one of the main bottlenecks in applications of machine learning, generating much interest in the problem of *active classification* where the learner is given an unlabeled pool of examples and her goal is to identify an accurate hypothesis using the minimum number of labels possible (Settles, 2011).

One of the most popular algorithmic paradigms is disagreement-based active classification (Hanneke et al., 2014). Under this approach, after observing k labels a version space $\mathcal{V}_k$ of the most promising classifiers is maintained, and the learner queries an example x if there are two hypotheses h_1 and h_2 belonging to $\mathcal{V}_k$ that disagree on the label of x. This approach has received much attention because it applies to generic hypothesis classes, it can be made robust to label noise, and it can be efficient by using a constrained cost-sensitive classification oracle, a problem for which there are many reasonable heuristics (Agarwal et al., 2018; Beygelzimer et al., 2010).

However, disagreement-based active classification suffers from two significant shortcomings. First, it queries uniformly any example on which there is disagreement even though intuitively some of these examples may be much more informative than others. Second, disagreement-based active classification algorithms tend to take a naive union bound over all hypotheses, which ignores many of the dependencies among the hypotheses. Indeed, recent work in pure exploration combinatorial and linear bandits has shown that such naive union bounds can be highly suboptimal and have a significant impact on empirical performance (Cao & Krishnamurthy, 2019; Jain & Jamieson, 2019; Katz-Samuels et al., 2020). Given that these naive union bounds are very loose and appear in the confidence bounds used by the algorithms, in practice, many works instead replace these union bounds with a constant that can be tuned to control the aggressiveness of the algorithm (Beygelzimer et al., 2010; Huang et al., 2015). Unfortunately, this constant introduces a hyperparameter to the active learning algorithm that is difficult to set before seeing lots of data.

We design a new algorithm for pool-based active classification that addresses these shortcomings. It optimizes a novel experimental design objective that finds the best

[1]University of Wisconsin, Madison, WI [2]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA. Correspondence to: Kevin Jamieson <jamieson@cs.washington.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

subset of examples in the disagreement region to query in order to identify the best classifier. It avoids wasteful union bounds by adapting to the geometry of the hypothesis space and thus avoiding the need to choose hyperparameters. We introduce a new notion of sample complexity inspired by experimental design that improves on disagreement-based active classification by a factor up to $\sqrt{n}$ where n is the size of the pool while being only a logarithmic factor worse than disagreement-based learning in the worst case.

1.1. Preliminaries

Let $\mathcal{X}$ denote the input space, and let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ denote a pool of examples. Let $\mathcal{H}$ denote a class of hypotheses where each $h : \mathcal{X} \mapsto \{0, 1\}$ assigns a label to each example in the pool. Let $\mathcal{H}_{\mathrm{x}} := \{(h(x_i))_{i\in[n]} : h \in \mathcal{H}\}$ denote the set of labelings over the pool induced by the hypothesis class $\mathcal{H}$. Let d denote the VC dimension of $\mathcal{H}$. When example $i \in [n]$ is queried, the agent receives label $Y_i \sim \mathrm{Bern}(\eta_i)$ where $\eta = (\eta_i)_{i=1}^n \in [0, 1]^n$. We define the error of a hypothesis $h \in \mathcal{H}$ on the pool of examples as given by

$$\begin{aligned}\mathrm{err}(h) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(Y_i \neq h(x_i)) \qquad (1)\\ &= \frac{1}{n}\sum_{i\in[n]} \eta_i(1 - h(x_i)) + (1 - \eta_i)h(x_i).\end{aligned}$$

Let $h_* := \arg\min_{h\in\mathcal{H}} \mathrm{err}(h)$ be the hypothesis of minimum error, and let $\nu = \mathrm{err}(h_*)$. The goal in active classification is to find an $h \in \mathcal{H}$ with error close to that of h_* using as few label queries as possible. In this paper, we quantify performance as follows:

Problem. Agnostic Pool Based PAC Active Classification: Given $\epsilon > 0, \delta \in (0, 1)$, identify an ϵ-good classifier, that is, an $h \in \mathcal{H}$ such that $\mathrm{err}(h) - \mathrm{err}(h_*) \leq \epsilon$ with probability at least $1 - \delta$ using as few labels as possible.

Remark 1. *The goal of finding an ϵ-good classifier over a pool of examples is closely related to the goal of using an active classification algorithm to find a classifier with good generalization. Suppose $VCdim(\mathcal{H}) = d$ and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$. For $i = 1, \ldots, n$ let $(x_i, y_i) \sim \mathcal{D}$. If $\widehat{h}$ satisfies $err(\widehat{h}) \leq \min_{h\in\mathcal{H}} err(h) + \epsilon$, then with probability at least $1 - \delta$*

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}(\widehat{h}(x) \neq y) \leq$$

$$\min_{h\in\mathcal{H}} \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) + \epsilon + O\Big(\sqrt{\frac{d\ln(1/\delta)}{n}}\Big).$$

by standard passive generalization bounds (Boucheron et al., 2005).

1.2. Main contributions

We briefly summarize our contributions:

- We cast pool-based active binary classification as an *adaptive experimental design problem* that computes an optimal sampling distribution over the pool of unlabelled examples. We demonstrate that an ϵ-good classifier can be obtained with probability at least $1 - \delta$ by requesting just $\gamma^*(\epsilon) + \rho^*(\epsilon)\log(1/\delta)$ labels if examples to label are drawn from the optimal design, where $\gamma^*(\epsilon)$ and $\rho^*(\epsilon)$ are problem-dependent quantities defined in the next section.
- Since this optimal design uses problem dependent information like η, it is not a constructive strategy or algorithm for a learner. Treating the sample complexity achieved by this optimal design as a target, we design an algorithm that performs sequential stages of experimental design to match the sample complexity of the optimal design, $\gamma^*(\epsilon) + \rho^*(\epsilon)\log(1/\delta)$ up to a $\log(1/\epsilon)$ factor. The algorithm employs the use of a novel estimator that appeals to a chaining argument. Unfortunately, the method is not computationally efficient.
- We propose a second algorithm that is computationally efficient given access to an empirical risk minimization oracle. The price for computational tractability is a slightly worse sample complexity. Besides being computationally efficient, our approach avoids the need to tune hyperparameters and the use of a *constrained* empirical risk minimization oracle which are required by other active learning algorithms (Beygelzimer et al., 2010; Huang et al., 2015).
- We compare our sample complexity results to those of state-of-the-art disagreement-based learning algorithms that are given in terms of the so-called disagreement coefficient. We demonstrate that our results, up to log factors, are never worse than previous results, but can be substantially better in certain cases.
- Empirically, we compare our procedure to state-of-the-art algorithms for the agnostic setting including variants of the importance weighted active learning algorithm (IWAL) (Beygelzimer et al., 2010) and active cover (Huang et al., 2015). We demonstrate that our method is superior across four image classification tasks.[1]

2. Experimental Design for Active Classification

We seek to identify an ϵ-good classifier by seeing as few labels as possible. To this end, we can take motivation from *experimental design* to consider the *optimal* sampling distribution over our pool of unlabeled examples $[n]$. For an arbitrary distribution $\lambda \in \triangle_n := \{p \in \mathbb{R}^n : p_i \geq 0, \forall i \in [n]; \sum_{i=1}^n p_i = 1\}$ suppose we sampled $I_1, \cdots, I_t \sim \lambda$ and then observed y_s for each $s \in [t]$. Then an unbiased natural estimator for the error of a classifier $h \in \mathcal{H}$ defined by (1)

[1] Code can be found at https://github.com/jifanz/ACED.

is given by

$$\widetilde{\text{err}}(h) = \frac{1}{t}\sum_{s=1}^{t}\frac{1/n}{\lambda_{I_s}}\mathbb{1}\{h(x_{I_s}) \neq y_s\}.$$

Indeed, by i.i.d. sampling from λ, we have for any $s \in [t]$

$$\begin{aligned}\mathbb{E}[\widetilde{\text{err}}(h))] &= \mathbb{E}\Big[\frac{1/n}{\lambda_{I_s}}\mathbb{1}\{h(x_{I_s}) \neq y_s\}\Big] \\ &= \sum_{i=1}^{n}\mathbb{P}(I_s = i)\frac{1/n}{\lambda_i}\mathbb{E}[\mathbb{1}\{h(x_i) \neq y_s\}|I_s = i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(Y_i \neq h(x_i)) = \text{err}(h)\end{aligned}$$

since by definition, $\mathbb{P}(I_s = i) = \lambda_i$. Likewise, an estimator for the excess risk is given by

$$\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h_*) = \frac{1}{t}\sum_{s=1}^{t}\frac{1/n}{\lambda_i}(\mathbb{1}\{h(x_{I_s}) \neq y_s\} - \mathbb{1}\{h_*(x_{I_s}) \neq y_s\}). \tag{2}$$

It is straightforward to show that the variance of $\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h_*)$ is upper bounded by $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i n^2}\mathbb{1}\{h_*(x_i) \neq h(x_i)\}$, using the upper bound $\mathbb{1}\{h(x_I) \neq y_s\} - \mathbb{1}\{h_*(x_I) \neq y_s\}) \leq \mathbb{1}\{h_*(x_I) \neq h(x_I)\}$. Applying Bernstein's inequality (and ignoring the $1/t$ term) with probability at least $1-\delta$

$$|\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h_*) - (\text{err}(h) - \text{err}(h_*))| \lessapprox \sqrt{\frac{\sum_{i=1}^{n}\frac{1}{\lambda_i n^2}\mathbb{1}\{h_*(x_i) \neq h(x_i)\}\log(|\mathcal{H}_x|/\delta)}{t}}. \tag{3}$$

This then suggests that to estimate the excess error of this particular h with probability at least $1-\delta$, it suffices to take t large enough to make the RHS of (3) less than ϵ. To upper bound the excess risk of every $h \in \mathcal{H}$ simultaneously, it suffices to take $t \geq \sup_{h\in\mathcal{H}}\frac{\sum_{i=1}^{n}\frac{1}{\lambda_i n^2}\mathbb{1}\{h_*(x_i)\neq h(x_i)\}}{\max\{\epsilon^2, (\text{err}(h)-\text{err}(h_*))^2\}}\log(|\mathcal{H}|/\delta)$. If we seek to *minimize* the total number of observations, we simply minimize over all $\lambda \in \triangle_n$, motivating the complexity measure:

$$\rho^*(\epsilon) := \inf_{\lambda\in\triangle_n}\sup_{h\in\mathcal{H}\setminus\{h_*\}}\frac{\sum_{i=1}^{n}\frac{1}{\lambda_i n^2}\mathbb{1}\{h_*(x_i) \neq h(x_i)\}}{\max(\text{err}(h) - \text{err}(h_*), \epsilon)^2}.$$

Thus, we'd expect that if $t \geq \rho^*(\epsilon)\log(|\mathcal{H}|/\delta)$ samples are drawn from the λ that minimizes $\rho^*(\epsilon)$, then $\widehat{h} = \arg\min_{h\in\mathcal{H}}\widetilde{\text{err}}(h)$ will be ϵ-good.

2.1. Sidestepping the Naive Union Bound

A significant shortcoming of the standard approach of applying Bernstein's inequality with a naive union bound is that the the naive union bound incurs an additional factor of $\log(|\mathcal{H}_x|)$ in the sample complexity. For infinite classes, $\log(|\mathcal{H}_x|)$ can be replaced by the VC-dimension of $\mathcal{H}_x$, however this can still be very loose. In practice, active learning algorithms replace $\log(|\mathcal{H}_x|)$ by a tunable parameter C_0 (Beygelzimer et al., 2010; Huang et al., 2015). Ideally C_0 would be chosen via cross-validation but since our data is being chosen adaptively, under an active algorithm that depends on C_0, it is unclear how to make the choice a priori.

To improve upon the naive union-bound we appeal to results from empirical process theory. Appealing to the Talagrand/Bousquet inequality (Boucheron et al., 2005), for all $h \in \mathcal{H}$, especially the empirical risk minimizer $\widehat{h} = \arg\min_{h\in\mathcal{H}}\widetilde{\text{err}}(h)$, we have

$$\begin{aligned}&\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h^*) - (\text{err}(h) - \text{err}\,h^*) \\ &\leq 2\mathbb{E}[\sup_{h\in\mathcal{H}}|\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h^*) - (\text{err}(h) - \text{err}\,h^*)|] \\ &\quad + \sqrt{\frac{\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\frac{1}{\lambda_i n^2}\mathbb{1}\{h_*(x_i) \neq h(x_i)\}\log(1/\delta)}{t}} \\ &\quad + \frac{4\sup_{i\in[n]}1/\lambda_i\log(1/\delta)}{3t}.\end{aligned}$$

Traditionally, we compute the expectation of the suprema using symmeterization to obtain the Rademacher complexity of $\mathcal{H}\setminus\{h_*\}$. In general, the Rademacher complexity is within a $\log(n)$ factor of the Gaussian Width (Bartlett & Mendelson, 2002). In particular,

$$\begin{aligned}&\mathbb{E}[\sup_{h\in\mathcal{H}}|\widetilde{\text{err}}(h) - \widetilde{\text{err}}(h^*) - (\text{err}(h) - \text{err}\,h^*)|] \\ &\leq \frac{1}{\sqrt{t}}\mathbb{E}_{\zeta\sim N(0,I)}\Big[\sup_{h\in\mathcal{H}}\sum_{i\in[n]}\frac{\zeta_i}{n\lambda^{1/2}}(h_*(x_i) - h(x_i))\Big].\end{aligned}$$

Using the same argument that motivated $\rho^*(\epsilon)$ but applying Bousquet's inequality instead of Bernstein's inequality, we introduce the following new complexity measure for active classification:

$$\gamma^*(\epsilon) := \inf_{\lambda\in\triangle_n}\mathbb{E}_\zeta\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{i\in[n]}\frac{\zeta_i(h_*(x_i)-h(x_i))}{n\lambda_i^{1/2}}}{\max(\text{err}(h) - \text{err}(h_*), \epsilon)}\Big]^2.$$

Analogous to above, if we ignore the $1/t$ term, we'd expect that if $t \geq \gamma^*(\epsilon) + \rho^*(\epsilon)\log(1/\delta)$ samples are drawn from the λ that minimizes the maximum of $\gamma^*(\epsilon)$ and $\rho^*(\epsilon)$, then $\widehat{h} = \arg\min_{h\in\mathcal{H}}\widetilde{\text{err}}(h)$ will be ϵ-good.

We can relate $\gamma^*(\epsilon)$ to $\rho^*(\epsilon)$ in the following way.

Proposition 1 (Katz-Samuels et al. (2020)). $\gamma^*(\epsilon) \leq c\log(|\mathcal{H}_x|)\rho^*(\epsilon) \leq cd\log(\frac{n}{d})\rho^*(\epsilon)$.

The first inequality parallels the application of Massart's finite class lemma to bound the Rademacher complexity in statistical learning theory and the second inequality follows

from the Sauer-Shelah Lemma. Katz-Samuels et al. (2020) also demonstrates a lower bound on $\gamma^*(\epsilon)$ that is dominated by $\rho^*(\epsilon)$. In the appendix, we show that $\gamma^*(\epsilon)$ matches the minimax rates for classification given for the hypothesis class of thresholds in (Castro & Nowak, 2008).

Main Takeaway: In Section 3 we will establish an algorithm that achieves a sample complexity of $(\gamma^*(\epsilon) + \rho^*(\epsilon)\log(1/\delta))\log(1/\epsilon)$ to obtain an ϵ-good classifier with probability greater than $1-\delta$. In the next section we compare this result to disagreement based methods. Note that we will write $\rho^* := \rho^*(0)$ and $\gamma^* := \gamma^*(0)$.

2.2. Comparison with the Disagreement Coefficient

To date, theoretically grounded active learning algorithms in the agnostic setting are disagreement region sampling methods. At the beginning of each round t these algorithms construct a version space $\mathcal{V} \subset \mathcal{H}$ which is defined to be the set of classifiers that have yet to be ruled out by the algorithm using the observed labels up to round $t-1$. These algorithms then choose x_{I_t} to be uniformly sampled from $\text{DIS}(\mathcal{V})$, the disagreement region, which is the set of points on which any two hypotheses in $\mathcal{V}$ disagree:

$$\text{DIS}(\mathcal{V}) = \{i : \exists h, h' \in \mathcal{V} \text{ s.t. } h(x_i) \neq h'(x_i)\}.$$

In the notation of the previous section, these algorithms are sampling from λ_t where λ_t is the uniform distribution supported on $\text{DIS}(\mathcal{V})$ (Hanneke et al., 2014).

The main complexity measure considered for disagreement based algorithms is the disagreement coefficient defined as

$$\theta(\xi) = \sup_{r \geq \xi} \frac{|\text{DIS}(B(h_*, r))|/n}{r}$$

where $B(h_*, r)$ is the ball of radius r centered at h_*:

$$B(h_*, r) = \{h \in \mathcal{H} : \frac{\sum_{i\in[n]} \mathbb{1}\{h_*(x_i) \neq h(x_i)\}}{n} \leq r\}.$$

We consider sample complexity results for finding an h with $\text{err}(h) \leq \nu + \epsilon$, where $\nu = \text{err}(h^*)$ under two common settings.

1. **The Agnostic Setting**: we make no assumptions on $\eta \in [0,1]^n$. In this case the best known sample complexities scale like
$$\theta(\epsilon)(\frac{\nu^2}{\epsilon^2} + \log(1/\epsilon))d$$
where d is the VC dimension of $\mathcal{H}$ (Hanneke et al., 2014). Note that the noiseless setting of $\eta \in \{0,1\}^n$ is a special case.
2. **The Tsybakov noise condition:** for some $a \in [1,\infty)$ and $\alpha \in (0,1]$ every $h \in \mathcal{H} \setminus \{h_*\}$ satisfies
$$\frac{\sum_{i\in[n]} \mathbb{1}\{h_*(x_i) \neq h(x_i)\}}{n} \leq a(\text{err}(h) - \text{err}(h_*))^\alpha.$$
In this case the best known sample complexities scale like:
$$a^2 \frac{1}{\epsilon^{2-2\alpha}} \theta(a\epsilon^\alpha) d \log(1/\epsilon).$$

We now compare our claimed sample complexity of $\gamma^*(\epsilon) + \rho^*(\epsilon)\log(1/\delta)$ to these known sample complexity results. Define $\Delta_{min} := \min_{h\in\mathcal{H}\setminus\{h_*\}} \text{err}(h) - \text{err}(h_*)$.

Proposition 2. • *Suppose that* $\eta \in \{0,1\}^n$.

$$\rho^*(\epsilon) \leq c \log(n\Delta_{min}^{-1} \vee \epsilon^{-1})\, \theta(\epsilon)[1 + \frac{\nu^2}{\epsilon^2}].$$

• *Suppose that the Tsabokov noise condition holds for some* $a \in [1,\infty)$ *and* $\alpha \in (0,1]$. *Then,*

$$\rho^*(\epsilon) \leq c a^2 \frac{1}{\epsilon^{2-2\alpha}} \theta(a\epsilon^\alpha) \log(n\Delta_{min}^{-1} \vee \epsilon^{-1}).$$

Recall that Propositions 1 shows $\gamma^*(\epsilon) \leq cd\rho^*(\epsilon)\log(n/d)$. Hence from Proposition 2, we see that our sample complexity, $\gamma^* + \rho^* \log(1/\delta)$ is always as good as the state-of-the-art sample complexities of disagreement-based learning up to logarithmic factors in n and ϵ^{-1} in both settings.

However, the converse is not true. In general the disagreement based active classification sample complexities can be substantially larger than ρ^* and γ^*.

Proposition 3. *There exists an instance where for sufficiently small* ξ, $\theta(\xi) \geq \Omega(n^{1/2})$ *while* $\rho^* = O(1)$ *and* $\gamma^* = \log(n)$.

We emphasize that this is not just a feature of the analysis; any algorithm that selects queries uniformly at random in the region of disagreement will perform poorly on the instance in the proposition. This gap demonstrates a provable improvement over prior art.

3. Fixed Confidence Algorithm

Algorithm 1 is an elimination-style algorithm, in the style of A^2 (Balcan et al., 2009; Dasgupta et al., 2007; Huang et al., 2015; Jain & Jamieson, 2019), but optimizes the querying distribution similarly to algorithms from the pure exploration linear bandits literature (Fiez et al., 2019; Katz-Samuels et al., 2020). It chooses a distribution λ_k over the examples in (4) that minimizes the confidence bounds from Theorem 1 and queries enough random examples from λ_k to ensure that the estimates of the difference in error rates, $\text{err}(h) - \text{err}(h_*)$, improve at least by a factor of 2 for all remaining hypotheses $h \in \mathcal{H}_k$. Using these improved estimates of the gaps, it then eliminates all hypotheses that can be shown to be suboptimal using the confidence bound in Theorem 1.

Given an estimator $\hat{\eta}$ for η, denote the induced estimate for

Algorithm 1 ACED (Active Classification using Experimental Design).

Input: Confidence level $\delta \in (0,1)$.
$\mathcal{H}_1 \longleftarrow \mathcal{H}$, $k \longleftarrow 1$, $\delta_k \longleftarrow \delta/2k^2$.
while $|\mathcal{H}_k| > 1$ **do**
Let λ_k and τ_k be the solution and value of the following optimization problem

$$\inf_{\lambda \in \triangle_n} \mathbb{E}_{\zeta \sim N(0,I)} \Big[\max_{h \in \mathcal{H}_k} \sum_{i \in [n]} h(x_i) \frac{\zeta_i}{n \lambda_i^{1/2}} \Big]^2 \tag{4}$$
$$+ 2 \log(\frac{1}{\delta_k}) \max_{h,h' \in \mathcal{H}_k} \max_{h,h' \in \mathcal{G}} \sum_{i=1}^n \frac{1}{\lambda_i n^2} \mathbb{1}\{h(x_i) \neq h'(x_i)\}$$

Set $N_k \longleftarrow c\tau_k 2^{2(k+1)}$ where c is a universal constant.
Query $I_1, \ldots, I_{N_k} \sim \lambda_k$ and receive rewards $y_1, \ldots, y_{N_k}$.
Let $\hat{\eta}_k := \hat{\eta}(\mathcal{H}_k, \delta_k)$ be the estimator defined in Theorem 1 for $\mathcal{H}_k$ with failure probability δ_k using the samples $\{(x_{I_s}, y_s)\}_{s=1}^{N_k}$
$\mathcal{H}_{k+1} \longleftarrow \mathcal{H}_k \setminus \{h \in \mathcal{H}_k : \exists h'$ such that $\widetilde{\text{err}}(h', \hat{\eta}_k) - \widetilde{\text{err}}(h, \hat{\eta}_k) + \frac{1}{2^{k+1}} \leq 0\}$.
$k \longleftarrow k+1$
end while
Return: $\mathcal{H}_k = \{\widehat{h}\}$.

the error as

$$\widetilde{\text{err}}(h, \hat{\eta}) = \frac{1}{n} \sum_{i \in [n]} \hat{\eta}_i (1 - h(x_i)) + (1 - \hat{\eta}_i) h(x_i).$$

Theorem 1. *Let $\mathcal{G} \subset \mathcal{H}$. There exists an estimator $\hat{\eta}(\mathcal{G}, \delta)$ for η constructed from t samples drawn i.i.d. from λ such that with probability at least $1 - \delta$,*

$$\sup_{h,h' \in \mathcal{G}} |[\widetilde{err}(h, \hat{\eta}) - \widetilde{err}(h', \hat{\eta})] - [err(h) - err(h')]|$$
$$\lesssim \sqrt{\frac{\log(2/\delta) \max_{h,h' \in \mathcal{G}} \sum_{i=1}^n \frac{1}{\lambda_i n^2} \mathbb{1}\{h(x_i) \neq h'(x_i)\}}{t}}$$
$$+ \sqrt{\frac{\mathbb{E}[\sup_{h \in \mathcal{G}} \sum_{i \in [n]} h(x_i) \frac{\zeta_i}{n \lambda_i^{1/2}}]^2}{t}}.$$

For now, we treat the estimator in Theorem 1 as a black-box and defer its discussion until Section 4.1. Note that unlike the Talagrand/Bousquet inequality presented before (3), *this confidence interval does not have a term depending on the inverse of the worst case importance weight*.

Algorithm 1 attains the following sample complexity.

Theorem 2. *Let $\delta \in (0,1)$ and $\epsilon > 0$. With probability at least $1 - \delta$ Algorithm 1 returns $\widehat{h} \in \mathcal{H}$ after τ samples where $err(\widehat{h}) \leq err(h_*) + \epsilon$ and*

$$\tau \lesssim \log(1/\epsilon)[\log(1/\delta)\rho^*(\epsilon) + \gamma^*(\epsilon)].$$

4. Fixed Budget Algorithm

Algorithm 2 Fixed Budget ACED.

Input: Budget T, tolerance $\epsilon > 0$
$N \longleftarrow \Big\lfloor T / \log_2(\epsilon^{-1}) \Big\rfloor$, and $\hat{\eta}_0 = 0$
for $k = 1, 2, \ldots, \Big\lfloor \log_2(\epsilon^{-1}) \Big\rfloor$ **do**
$\tilde{h}_k \longleftarrow \arg\min_{h \in \mathcal{H}} \widetilde{\text{err}}(h, \hat{\eta}_{k-1})$.
Let λ_k be the solution of the following optimization problem

$$\inf_{\lambda \in \triangle_n} \mathbb{E}_{\zeta \sim N(0,I)} \Big[\max_{h \in \mathcal{H}} \frac{\sum_{i \in [n]} (\tilde{h}_k(x_i) - h(x_i)) \frac{\zeta_i}{n \lambda_i^{1/2}}}{2^{-k+1} + \widetilde{\text{err}}(h, \hat{\eta}_{k-1}) - \widetilde{\text{err}}(\tilde{h}_k, \hat{\eta}_{k-1})} \Big] \tag{5}$$

Sample $\{x_{I_1}, \ldots, x_{I_N}\} \sim \lambda_k$.
Query $x_{I_1}, \ldots, x_{I_N}$ and observe $y_1, \ldots, y_N$.
Compute an estimate $\hat{\eta}_k$.
end for
Return: $\arg\min_{h \in \mathcal{H}} \widetilde{\text{err}}(h, \hat{\eta}_k)$

In many applications, the agent is given a budget of T queries and a performance target $\epsilon > 0$, and the goal is to maximize the probability of outputing a classifier $\widehat{h} \in \mathcal{H}$ such that $\text{err}(\widehat{h}) \leq \text{err}(h_*) + \epsilon$. We design a new algorithm for this setting that can be made computationally efficient given access to a weighted classification oracle (defined shortly).

Algorithm 2 splits the budget into $\Big\lfloor \log(\epsilon^{-1}) \Big\rfloor$ phases. In each phase, the algorithm computes the design that optimizes (5), the objective of which approximates $\mathbb{E}\Big[\max_{h \in \mathcal{H} \setminus \{h_*\}} \frac{\sum_{i \in [n]} \frac{\zeta_i}{n \lambda^{1/2}} (h_*(x_i) - h(x_i))}{\max(\text{err}(h) - \text{err}(h_*), 2^{-k+1})} \Big]^2$. The algorithm can use any estimator $\hat{\eta}_k$ at each round k. The next theorem uses the estimator of Theorem 1.

Theorem 3. *Let $T \in \mathbb{N}$ and $\epsilon > 0$. Let $\widehat{h}$ denote the $h \in \mathcal{H}$ returned by Algorithm 2. There exists an estimator $\hat{\eta}_k$ using the samples $\{(x_{I_s}, y_s)\}_{s=1}^N$ in round k of Algorithm 2 such that for an absolute constant $c > 0$*

$$\mathbb{P}(err(\widehat{h}) \geq err(h_*) + \epsilon)$$
$$\leq \log(n\epsilon^{-1})^2 \exp\Big(- \frac{cT}{\log(\epsilon^{-1})[\gamma^*(\epsilon) + \rho^*(\epsilon)]} \Big).$$

If $T \geq c \log(\log(\epsilon^{-1})) \log(1/\delta) \log(\epsilon^{-1})[\gamma^*(\epsilon) + \rho^*(\epsilon)]$, then with probability at least $1 - \delta$, Algorithm 2 outputs $\widehat{h} \in \mathcal{H}$ such that $\text{err}(\widehat{h}) \leq \text{err}(h_*) + \epsilon$. The proof of Theorem 3 leverages the estimator defined in Theorem 1 for $\mathcal{H}$ and failure probability $\delta_k = \exp(-\Theta(N/\gamma_k))$ with γ_k equal to the value of (5).

Remark 2. *Given $\{(I_t, y_t)\}_{t=1}^T$ where $I_t \sim \lambda$ define*

$$\hat{\eta}_\gamma^{(Importance)} = \frac{1}{T} \sum_{t=1}^T \frac{y_t}{\lambda_{I_t} + \gamma} \mathbf{e}_{I_t}. \tag{6}$$

If importance-weighted estimator $\hat{\eta}_\gamma^{(Importance)}$ with $\gamma = 0$ is used in Algorithm 2 (with a slightly modified objective func-

tion in (5), see the Supplementary Material), one can obtain a computationally efficient algorithm whose probability of error scales as

$$\mathbb{P}(err(\widehat{h}) \geq err(h_*) + \epsilon) \leq \log(n\epsilon^{-1})^2 \exp(-\frac{T - \log(|\mathcal{H}_{\mathrm{x}}|)\psi^*(\epsilon)}{\log(\epsilon^{-1})[\gamma^*(\epsilon) + \rho^*(\epsilon) + \psi^*(\epsilon)]})$$

where

$$\psi^*(\epsilon) := \min_{\lambda \in \triangle_n} \max_{\substack{i \in [n]: \exists h \in \mathcal{H} \\ h_*(x_i) \neq h(x_i)}} \frac{1/n\lambda_i}{\max(\epsilon, err(h) - err(h_*))}.$$

There are instances where $\psi^*(\epsilon) \gg \gamma^*(\epsilon)$ and therefore the cost of computational efficiency is a worse sample complexity. See the appendix for more details.

4.1. Discussion of Theorem 1

Theorem 1 above demonstrates the existence of an estimator that avoids any dependence on $\log(|\mathcal{H}_{\mathrm{x}}|)$. The construction of the estimator in Theorem 1 uses generic chaining, a technique that builds a highly optimized union bound to avoid extraneous logarithmic factors (Talagrand, 2014). Generic chaining is most easily applied when a given estimator $\hat{\eta}$ satisfies the property that $\widetilde{\mathrm{err}}(h, \hat{\eta}) - \widetilde{\mathrm{err}}(h', \hat{\eta})$ is sub-Gaussian for every "direction" $h - h'$ of interest (e.g., see (Katz-Samuels et al., 2020)). Though the $\hat{\eta}_\gamma^{(\text{Importance})}$ estimator has sub-Gamma tails in general, ruling out its use, the following result shows that for $h - h'$ in a ball under a certain norm, we can construct an estimator for $\widetilde{\mathrm{err}}(h, \hat{\eta}) - \widetilde{\mathrm{err}}(h', \hat{\eta})$ with a sub-Gaussian-like tail.

Proposition 4. *Fix $\lambda \in \triangle_n$, $\delta \in (0,1)$, and $h, h' \in \mathcal{H}$. If T samples are taken from λ and $\hat{\eta} := \hat{\eta}_\gamma^{(\text{Importance})}$ is computed with $\gamma = \sqrt{\frac{\log(2/\delta)}{3\sum_{i=1}^n \frac{1}{\lambda_i n^2}\mathbb{1}\{h(x_i) \neq h'(x_i)\}}}$ then with probability at least $1 - \delta$*

$$|[\widetilde{err}(h, \hat{\eta}) - \widetilde{err}(h', \hat{\eta})] - [err(h) - err(h')]| \leq \left(\sqrt{\tfrac{2}{3}} + 1\right)\sqrt{\frac{2\sum_{i=1}^n \frac{1}{\lambda_i n^2}\mathbb{1}\{h(x_i) \neq h'(x_i)\}\log(\frac{2}{\delta})}{t}}.$$

The idea behind Theorem 1 is to apply generic chaining to all $h - h'$, but to use a different $\hat{\eta}$ (specifically, a different γ) based on the size of $h - h'$ prescribed by Proposition 4. Details of the technique can be found in the supplementary materials.

4.2. Computationally Efficient Experimental Design

In this section, we discuss how to solve (5) efficiently given access to a weighted empirical risk minimization oracle, which we will introduce shortly. First, note that minimizing (5) is equivalent to minimizing $\mathbb{E}_{\zeta \sim N(0,I)}[\max_{h \in \mathcal{H}} f(\lambda; h; \zeta)]$ with respect to λ where

$$f(\lambda; h; \zeta) := \frac{\sum_{i \in [n]}(\tilde{h}_k(x_i) - h(x_i))\frac{\zeta_i}{n\lambda_i^{1/2}}}{2^{-k+1} + \widetilde{\mathrm{err}}(h, \hat{\eta}_{k-1}) - \widetilde{\mathrm{err}}(\tilde{h}_k, \hat{\eta}_{k-1})} := \frac{\sum_{i \in [n]}(\tilde{h}_k(x_i) - h(x_i))\frac{\zeta_i}{n\lambda_i^{1/2}}}{2^{-k+1} + \sum_{i \in [n]}(1 - 2\hat{\eta}_{k-1,i})(\tilde{h}_k(x_i) - h(x_i))}.$$

It is known that $\mathbb{E}_{\zeta \sim N(0,I)}[\max_{h \in \mathcal{H}} f(\lambda; h; \zeta)]$ is convex in λ (Katz-Samuels et al., 2020), hence we perform the minimization over λ via stochastic mirror descent with stochastic gradient $g(\lambda, \zeta) = \nabla f(\lambda, \widetilde{h}; \zeta)$ where $\zeta \sim \mathcal{N}(0, I)$ and $\widetilde{h} \in \arg\max_{h \in \mathcal{H}} f(\lambda, h; \zeta)$. To obtain $\widetilde{h}$ for a fixed λ and ζ, first note that the value $\max_{h \in \mathcal{H}} f(\lambda; h; \zeta)$ is equal to

$$\min_{r \in \mathbb{R}^+} r \text{ subject to } ar + b + \max_{h \in \mathcal{H}} \sum_{i \in [n]} (c_i r + d_i)h(x_i) \leq 0$$

where $a = -2^{-k+1} - \sum_{i \in [n]}(1 - 2\hat{\eta}_{k,i})\tilde{h}_k(x_i)$, $b = \sum_{i \in [n]} \frac{\zeta}{n\lambda_i^{1/2}}\tilde{h}_k(x_i)$, $c_i = 1 - 2\hat{\eta}_{k-1,i}$ and $d_i = -\frac{\zeta}{n\lambda_i^{1/2}}$.

For any fixed positive value of r it suffices to check the constraint. We can then use a line search procedure to find the minimizing value of r (details in Appendix K).

Thus we have reduced to checking the constraint for a fixed $r \in \mathbb{R}^+$. Specifically, the difficulty is to solve for $\max_{h \in \mathcal{H}} \sum_{i \in [n]} w_i \cdot h(x_i)$ where w_i are arbitrary weights. This can be reduced to weighted $0/1$-loss minimization problem that is solvable by a weighted classification oracle:

$$\text{oracle}(\{\tilde{w}_i, \tilde{x}_i, \tilde{y}_i\}_{i=1}^n) := \arg\min_{h \in \mathcal{H}} \sum_{i \in [n]} \tilde{w}_i \cdot \mathbf{1}\{h(\tilde{x}_i) \neq \tilde{y}_i\}$$

for inputs $\{\tilde{w}_i, \tilde{x}_i, \tilde{y}_i\}_{i=1}^n$. Then,

$$\max_{h \in \mathcal{H}} \sum_{i \in [n]} w_i \cdot h(x_i) = \text{oracle}(\{|w_i|, x_i, \mathbf{1}\{w_i \geq 0\}\}_{i=1}^n).$$

5. Implementation and Experiments

In the previous section we reduced the experimental design objective of (5) to a weighted 0/1 loss classification problem using weights that are functions of the estimated vector $\widehat{\eta}$. In practice we replace this 0/1 loss with a surrogate convex loss, namely the logistic loss. However, to implement Algorithm 2 we still have to specify the choice of estimator $\hat{\eta}$. Though the estimator specified in Theorem 1 is theoretically grounded, it is difficult to implement in practice since it involves a costly constrained linear optimization problem over the set of hypothesis in $\mathcal{H}_k$. As described in Remark 2, it is still possible to have a theoretical guarantee for other estimators such as the IPS estimator. As described precisely in Appendix K, in our implementation we take the estimate for $\hat{\eta}_k$ to be

$$\left[\hat{\eta}_k^{(\text{Naive})}\right]_i = \text{average}(\{y_s^{(j)} : I_s^{(j)} = i, s \in [N_j], j \in [k]\}),$$

i.e. a simple average of the labels we see. Here $I_s^{(j)}$ indexes the s-th query we made in round j. In our experiments we only considered the persistent noise setting (i.e., querying the same image more than once would always return the same label as before, or formally, $\eta_i \in \{0, 1\}$). Thus, if we sample a point $x_{I_s}^{(j)}$ (i.e., $I_s^{(j)}$) more than once, we set $y_s^{(j)}$ to be the previously observed label and we did not count this observation in our count of total labels taken. To take advantage of all of the labels observed so far, we also employ a water-filling technique for sampling in practice (details in Appendix K).

Baselines. To validate Algorithm 2 we conducted a set of experiments against the following baselines that are considered to be state-of-the-art theoretically-justified methods in disagreement based active learning. Our set of methods are chosen based on the ones considered in (Huang et al., 2015), the most recent work of relevance. Details on the precise implementations of these methods are available in the supplementary materials in Appendix K.

- Passive: We considered a passive baseline where we uniformly at random choose samples from our pool, retrain our model on our current samples and report the accuracy.
- Importance Weighted Active Learning (IWAL) : IWAL was originally introduced in Beygelzimer et al. (2009) and is an active learning algorithm in the streaming setting. Our implementation is based on the algorithm presented in Beygelzimer et al. (2010) which we refer to as IWAL0. We also consider variants, IWAL1, and oracular versions ORA-IWAL0, ORA-IWAL1 detailed in Huang et al. (2015).
- Online Active Cover (OAC): OAC is described in Huang et al. (2015). We used the implementation of OAC that is available in Vowpal Wabbit (Agarwal et al., 2014).

Datasets. We evaluate on the following four real datasets.

- **MNIST 0-4 vs 5-9** (LeCun et al., 1998). We considered the standard MNIST dataset but in a binary setting where digits 0-4 are labelled as 0, and 5-9 are labelled as 1. Our pool has 50000 images in total, and we classified based on the flattened images (784 dimensions).
- **SVHN 2 vs 7** (Netzer et al., 2011). We considered the binary classification problem of determining whether a digit was a 2 or a 7 (ignoring all other images). To prevent the logistic classifier from overfitting to arbitrary labels and to restrict the hypothesis class $\mathcal{H}$, we downsample the images to 512 dimensional feature vectors through PCA. There are 16180 images in total.
- **CIFAR Bird vs Plane** (Netzer et al., 2011). We considered the binary classification problem of determining whether a digit was a bird or a plane (ignoring all other images). To prevent the logistic classifier from overfitting to arbitrary labels and to restrict the hypothesis class $\mathcal{H}$, we downsample the images to 576 dimensional feature vectors through PCA. There are 10000 images in total.
- **FashionMNIST T-shirt vs Pants** (Xiao et al., 2017). We considered the binary classification problem of T-shirt vs Pants. Our pool has 12000 images in total, and we classified based on the flattened images (784 dimensions).

Implementations. We use two implementations to measure the performances of the algorithms.

- Implementation from Vowpal Wabbit (Agarwal et al., 2014) that is used by Huang et al. (2015). The implementation employs an online learner that only updates based on the latest queried label, therefore has time complexity that scales linearly in the number of images n.
- For our implementation in a batched setting, we retrain the entire classifier to convergence every time new labels become available. We find that the online learner of above can perform significantly better than our batched learner during the first few batches of training. However, our implementation has more stable accuracies during the course of training and performs slightly superior ($< 1\%$) in final accuracy. This comes at a cost of an $O(n^2)$ time complexity, which is too expensive in some of the settings.

In particular, we only use the Vowpal Wabbit implementation for the OAC experiments and the oracular variants of IWAL algorithms for our MNIST experiement, due to the high computation cost for running these algorithms with exhaustive hyperparameter search. However, we think this is still a fair comparison when evaluating some baselines using the the two implementations since it is the best one can achieve for those baselines within a computation budget (single machine with state-of-art commercialized CPU that runs for a month).

Hypothesis Class. In our implementation, we took the hypothesis space to be the set of linear separators in the underlying feature space. We used the logistic regression implementation in Scikit-learn (Pedregosa et al., 2011) for our underlying classification oracle.

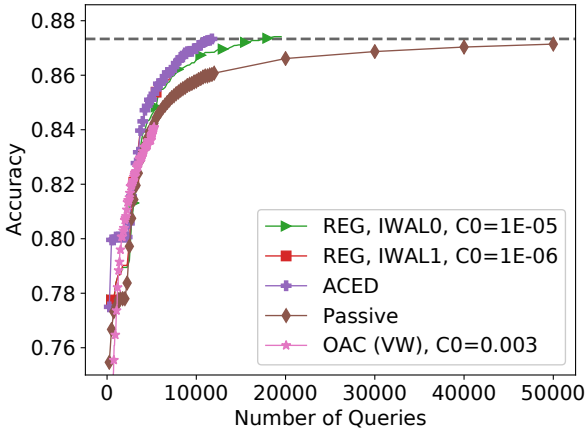

Figure 1. MNIST Performance

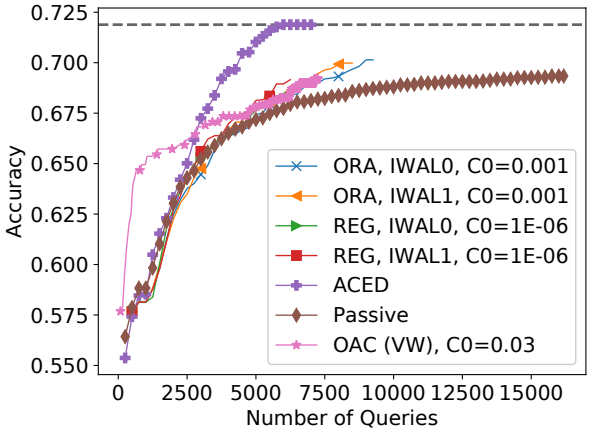

Figure 2. SVHN Performance

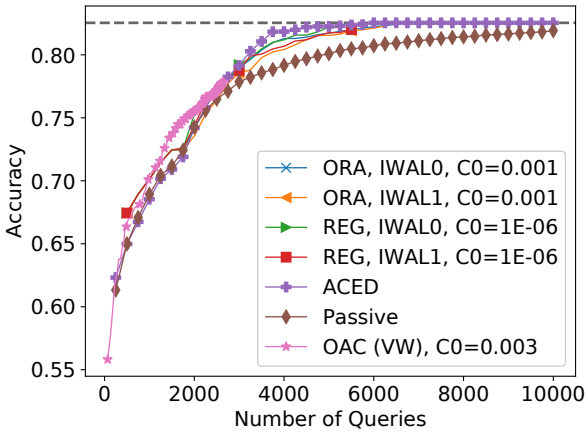

Figure 3. CIFAR Performance

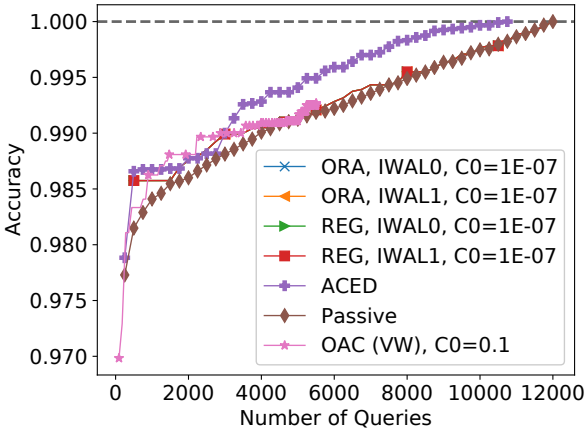

Figure 4. FashionMNIST

Discussion. For each of the binary classification datasets, we plot the running maximum accuracy on the unlabelled pool against the number of queries taken as in Figure 1,2,4,3 (full scale images included in Appendix L). The passive curves are evaluated based on the averages of 10 runs. In the CIFAR experiment, ACED is an average over 5 runs. We find the curve in this setting to be very consistent, and that the standard deviations are minimal for visualization. All of the other curves are evaluated based on a single run. For baselines algorithms proposed in the streaming setting (variants of IWAL and OAC), in each round we uniformly sample an example from the pool, and feed a fixed number of passes. We select the best C_0 based on which hyperparameter setting takes the least amount of queries to reach the same level of accuracy. Detailed hyperparameters considered for the baselines are included in Appendix M. Furthermore, to demonstrate active gains in generalization, we include plots on holdout test sets in Appendix N.

On all four datasets, our algorithm outperforms other baselines by taking much fewer queries to reach the passive accuracy on the entire dataset. Sometimes the active learning algorithms even beat the passive accuracy on the whole dataset, which is a known phenomenon of active learning studied by Mussmann & Liang (2018). For the MNIST dataset, we do not include performance curves for the oracular variants of IWAL, since the Vowpal Wabbit implementation turns out to be performing at random chance. We also notice that OAC stops taking queries very early on (no longer making queries when given more passes over the pool). However, when increasing C_0, the aggressiveness to make a query, OAC starts performing worst than passive pretty easily. We include Figure 9 in the appendix to demonstrate how sensitive the OAC curves are to the hyperparameter C_0, which one cannot tune in real applications.

As a special case, on the FashionMNIST dataset, our binary classification task is linearly separable and the baseline

methods fail miserably. For all of the IWAL algorithms on this dataset, we searched in an extended range of hyperparameters than the ones used in the other three tasks. When fixing the random order of the stream, however, all of the baselines become equivalent, and perform almost identical to passive. Since in practice, only one set of hyperparameters can be deployed, this again demonstrates the shortcoming of these baseline algorithms, whereas our method does not rely on any aggressiveness hyperparameter.

6. Related Work and Discussion

Active Classification: Active classification has received much attention with a large number of theoretical and empirical works (see (Hanneke et al., 2014) and (Settles, 2011) for excellent surveys). Cohn et al. (1994) initiated research into the study of disagreement based active classification algorithms, proposing CAL for the realizable setting. Balcan et al. (2009) extended disagreement-based active classification to the agnostic case, introducing the method, A^2. Hanneke (2007) provided a general analysis of A^2 in terms of the disagreement coefficient, with follow-up works improving on the sample complexity of this approach (Dasgupta et al., 2007; Hanneke, 2009; Hanneke et al., 2011; Koltchinskii, 2010; Hanneke et al., 2014). The results in Section 2.2 show that our sample complexities are never worse than the ones obtained by this line of work.

An extension of this line of work has aimed to attain similar sample complexities, while leveraging an empirical risk minimization oracle to design more practical algorithms (Dasgupta et al., 2007; Hsu, 2010; Beygelzimer et al., 2010; Huang et al., 2015). With the exception of Huang et al. (2015), these methods tend to have a conservative query policy that samples uniformly in the disagreement region, leading to an onerous label requirement. While Huang et al. (2015) has a more aggressive query policy that does not sample uniformly in the disagreement region, their sample complexity result could also be obtained by sampling uniformly in the disagreement region and, therefore, their theoretical result does not reflect gains from a careful selection of points in the disagreement region. In particular, the dominant term is still the disagreement coefficient and, hence, it can be much worse than our sample complexity on instances such as the one in Proposition 3.

Recently, Jain & Jamieson (2019) showed that active classification in the pool-based setting is an instance of combinatorial bandits, an observation that is central to our analysis. They provided the first analysis that shows the contribution of each example to the sample complexity providing a more fine-grained result than the disagreement coefficient. We improve on this work by optimizing the sampling distribution in the region of disagreement and using improved estimators such as the one in Theorem 1. Proposition 4 of Katz-Samuels et al. (2020) implies that our sample complexity is always better than the sample complexity in Jain & Jamieson (2019).

Finally, we also note that Zhang & Chaudhuri (2014) also give an algorithm that improves on disagreement-based active learning, but the sample complexity of their algorithm is difficult to interpret and their algorithm is not computationally efficient.

Linear and Combinatorial Bandits. ρ^* has been shown to be the dominant term in a lower bound for pure exploration linear bandits and combinatorial bandits (Soare et al., 2014; Chen et al., 2017; Fiez et al., 2019). Recently Katz-Samuels et al. (2020) introduced the notion of γ^* for linear and combinatorial bandits, showing that it is a lower bound for any non-interactive oracle MLE algorithm. One of our contributions is making the connection between the active classification and linear/combinatorial bandit literature, and showing that we can leverage the results from this work to obtain improved sample complexities for agnostic active classification.

Acknowledgements

The authors would like to thank Tzu-Kuo Huang, Alekh Agarwal and John Langford for their help with Vowpal Wabbit baseline experiments. Computational resources from Amazon Web Services were generously gifted as part of an Amazon Research Award. The work of KJ is supported in part by grants NSF RI 1907907 and NSF CCF 2007036.

References

Agarwal, A., Chapelle, O., Dudík, M., and Langford, J. A reliable effective terascale linear learning system. *The Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

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

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56, 2009.

Beygelzimer, A., Hsu, D. J., Langford, J., and Zhang, T. Agnostic active learning without constraints. In *Advances in neural information processing systems*, pp. 199–207, 2010.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Cao, T. and Krishnamurthy, A. Disagreement-based combinatorial pure exploration: Sample complexity bounds and an efficient algorithm. In *Conference on Learning Theory*, pp. 558–588, 2019.

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

Chen, L., Gupta, A., Li, J., Qiao, M., and Wang, R. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pp. 482–534, 2017.

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

Dasgupta, S., Hsu, D. J., and Monteleoni, C. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 20:353–360, 2007.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pp. 10666–10676, 2019.

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

Hanneke, S. Adaptive rates of convergence in active learning. In *COLT*. Citeseer, 2009.

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

Hanneke, S. et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Hsu, D. J. *Algorithms for active learning*. PhD thesis, UC San Diego, 2010.

Huang, T.-K., Agarwal, A., Hsu, D. J., Langford, J., and Schapire, R. E. Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems*, pp. 2755–2763, 2015.

Jain, L. and Jamieson, K. G. A new perspective on pool-based active classification and false-discovery control. In *Advances in Neural Information Processing Systems*, pp. 13992–14003, 2019.

Karampatziakis, N. and Langford, J. Online importance weight aware updates. *arXiv preprint arXiv:1011.1576*, 2010.

Katz-Samuels, J., Jain, L., Karnin, Z., and Jamieson, K. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *arXiv preprint arXiv:2006.11685*, 2020.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ledoux, M. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. ISBN 9780821837924. URL https://books.google.com/books?id=mCX_cWL6rqwC.

Mussmann, S. and Liang, P. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. *arXiv preprint arXiv:1812.01815*, 2018.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Settles, B. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pp. 1–18, 2011.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Talagrand, M. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Vershynin, R. *High-Dimensional Probability*. 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, C. and Chaudhuri, K. Beyond disagreement-based agnostic active learning. *Advances in Neural Information Processing Systems*, 27:442–450, 2014.