Active Learning with Disagreement Graphs

Corinna Cortes [1] **Giulia DeSalvo** [1] **Claudio Gentile** [1] **Mehryar Mohri** [1 2] **Ningshan Zhang** [3]

Abstract

We present two novel enhancements of an on-line importance-weighted active learning algorithm IWAL, using the properties of disagreements among hypotheses. The first enhancement, IWAL-D, prunes the hypothesis set with a more aggressive strategy based on the disagreement graph. We show that IWAL-D improves the generalization performance and the label complexity of the original IWAL, and quantify the improvement in terms of a *disagreement graph coefficient*. The second enhancement, IZOOM, further improves IWAL-D by adaptively zooming into the current version space and thus reducing the best-in-class error. We show that IZOOM admits favorable theoretical guarantees with the changing hypothesis set. We report experimental results on multiple datasets and demonstrate that the proposed algorithms achieve better test performances than IWAL given the same amount of labeling budget.

1. Introduction

In standard supervised learning, often a significant amount of labeled data is needed to learn an accurate predictor. But, while unlabeled data is virtually unlimited and available at no cost in many applications such as natural language processing or image recognition, labeled data is typically very costly, since it requires human inspection, often by domain experts. This challenge of learning with a limited labeling budget, or, equivalently, while minimizing the number of labels requested, motivates the scenario of *active learning*.

The goal of active learning algorithms is to learn an accurate predictor with as few labels as possible. There are two standard settings. In the so-called *pool setting*, the learner is provided with a pool of i.i.d. unlabeled points, from which he interactively selects points to label. In the *on-line setting*, the learner receives a sequence of i.i.d. unlabeled points and, at each round, decides on whether to request the label of the current point. In both settings, once the labeling budget is reached, the learner returns a predictor chosen out of a hypothesis set, which is hoped to admit a smaller generalization error than if the labeling budget had been spent at random. An active learning algorithm for the on-line setting can also be applied to the pool setting, since the learner can stream the pool of samples.

Active learning algorithms are typically evaluated in terms of their label complexity, that is how many labels are sufficient to obtain a small generalization error, ϵ, with high probability. In the on-line setting, (Cohn et al., 1994) proposed a disagreement-based algorithm, CAL, with label complexity $\log(\frac{1}{\epsilon})$, when the data is separable, as opposed to random sampling with a label complexity of $\frac{1}{\epsilon}$. The CAL algorithm maintains a version space, which includes all the classifiers that are consistent with the observed labels, and requests labels only when there is some disagreement among the predictors in the current version space.

Following the idea of disagreement-based active learning, several algorithms extended CAL to the non-separable case: A^2 (Balcan et al., 2006; Hanneke, 2007), DHM (Dasgupta et al., 2008), IWAL (Beygelzimer et al., 2009; 2010), which admit theoretical guarantees in generalization error and label complexity. In particular, the label complexities of these disagreement-based algorithms are bounded in terms of two critically quantities: the loss of the best predictor h^* in the hypothesis set, and the so-called *disagreement coefficient* (Hanneke, 2007) which depends on the disagreement between h^* and other hypotheses in the neighborhood of h^*. More recently, Zhang and Chaudhuri (2014) moved beyond disagreement-based active learning and instead combined confidence-rated predictors with disagreements to improve the label complexity. Besides disagreement-based active learning, another line of work (Dasgupta et al., 2005; Balcan et al., 2007; Balcan and Long, 2013; Awasthi et al., 2014; 2015; Zhang, 2018) studied learning linear separators by labeling samples close to the current estimate of the decision boundary. This type of algorithms admit favorable label complexity assuming the uniform distribution over the unit sphere or a log-concave distribution. Cortes et al. (2019) recently proposed a region-based active learn-

[1]Google Research, New York, NY, USA; [2]Courant Institute, New York, NY, USA; [3]Leonard N. Stern School of Business, New York University, New York, NY, USA. Correspondence to: Ningshan Zhang <nzhang@stern.nyu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

ing algorithm, where a favorable partition of the input space into sub-regions is assumed, and the algorithm optimally allocates unlabeled samples to active learning algorithms operating on each region.

Although many active learning algorithms are based on the notion of disagreement between hypotheses at each round, we are not aware of any making use of the average disagreements between all hypotheses, which are quantities that can be accurately estimated without requiring labels. More generally, we will talk about a *disagreement graph* by referring to the fully connected graph whose vertices are the hypotheses and whose edges are labeled with the average disagreements between the vertices. One key idea in this paper is to make use of the disagreement graph to improve an existing disagreement-based active learning algorithm.

Clearly, depending on the distribution, the graph may be more or less favorable. One favorable scenario is where the best-in-class vertex is within an isolated cluster of vertices, and the average disagreements within this cluster is small. In that case, an active learning algorithm will be able to quickly locate this cluster, and then identify the best-in-class vertex while requesting only a few labels since the disagreements are small. In this paper, we propose an active learning algorithm using the disagreement graph, and give guarantees in terms of properties of the disagreement graph, which measures how favorable the graph will be. While the learning bound is distribution-independent, the quality of the disagreement graph does depend on the distribution.

Another critical quantity that determines the performance of active learning algorithms is the loss of the best-in-class predictor: a smaller best-in-class error helps active learning algorithms achieve high accuracy with fewer labels. For most active learning algorithms with theoretical guarantees, the hypothesis set and therefore the best-in-class error are fixed. Thus, the second key contribution of this paper is to reduce the best-in-class error by adaptively enriching the original hypothesis set near the current best one, while running an active learning algorithm. The challenge for doing so is that the standard theoretical guarantees for existing active learning algorithm do not hold for a changing hypothesis set. Nevertheless, we will show that, by exploiting a key property of the disagreements, theoretical guarantees can be proven for the generalization error and the label complexity, both depending on a smaller best-in-class error.

The rest of this paper is organized as follows. In Section 2, we introduce the notation relevant to our analysis and define the learning scenario. In Section 3, we briefly describe the IWAL algorithm (Beygelzimer et al., 2009), which serves as an important baseline. In Section 4, we present a refined analysis of the label complexity bound of IWAL. In Section 5, we devise a new hypothesis pruning strategy enhancing the original strategy of IWAL, by using the disagreement graph. In Section 6, we further improve the above disagreement-graph-based IWAL by adaptively *zooming* into the function space to decrease the best-in-class error. We prove favorable label complexities and generalization bounds for our new algorithms using the properties of the disagreement graph. Finally, we report the results of several experiments demonstrating the favorable empirical performance of our new algorithms in Section 7.

2. Preliminaries

In this section, we introduce the relevant notation and describe the active learning scenario we examine.

We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ the input space and by $\mathcal{Y} = \{-1, +1\}$ the binary output space. We assume that training and test points are drawn i.i.d. according to an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We will denote by $\mathcal{D}_{\mathcal{X}}$ the marginal distribution induced by $\mathcal{D}$ over the input space $\mathcal{X}$.

Let $\mathcal{H}$ be the hypothesis set of functions mapping from $\mathcal{X}$ to $\mathcal{Z} \subseteq \mathbb{R}$. Let $\ell \colon \mathcal{Z} \times \mathcal{Y} \to [0, 1]$ be a loss function. For any hypothesis h, we will denote by $R(h)$ the generalization error or expected loss of h: $R(h) = \mathbb{E}[\ell(h(x), y)]$. We denote by $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$ the best-in-class hypothesis in $\mathcal{H}$ and by $R^* = R(h^*)$ its expected loss.

We also denote by $\mathcal{L}(h(x), h'(x))$ the maximum disagreement value between two hypotheses h, h' on point $x \in \mathcal{X}$:

$$\mathcal{L}(h(x), h'(x)) = \max_{y \in \mathcal{Y}} \big|\ell(h(x), y) - \ell(h'(x), y)\big|.$$

With a slight abuse of notation, we denote by $\mathcal{L}(h, h') = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathcal{L}(h(x), h'(x))]$ the expected maximum disagreement value between h and h' over $\mathcal{D}_{\mathcal{X}}$.

We consider the *on-line active learning scenario* where, at each round $t \in [T] = \{1, \ldots, T\}$, the learner receives an input point $x_t \in \mathcal{X}$ drawn i.i.d. according to $\mathcal{D}_{\mathcal{X}}$ and either requests its label, in which case she receives its label y_t, or simply chooses not to solicit its label.

The quality of an active learning algorithm is measured by two quantities in this setting: the generalization error of the hypothesis $\widehat{h}_T$ returned by the algorithm after T rounds, and the total number of labels requested within the T rounds.

3. Background on the IWAL Algorithm

Here, we give a brief description of the importance-weighted active learning algorithm of Beygelzimer et al. (2009), IWAL, which we extend and improve upon in later sections. We will also indicate the guarantees previously shown for this algorithm. We consider the IWAL algorithm since it can be used with different loss functions and hypothesis sets, and is agnostic to noise levels.

Given a finite hypothesis set $\mathcal{H}$, IWAL operates on a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ drawn i.i.d. according to $\mathcal{D}$. The algorithm maintains at any time t a version space $\mathcal{H}_t \subseteq \mathcal{H}$, with $\mathcal{H}_1 = \mathcal{H}$. At time t, the algorithm flips a coin $Q_t \in \{0, 1\}$ with bias p_t defined by

$$p_t = \max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x_t), g(x_t)). \tag{1}$$

If $Q_t = 1$, IWAL requests the label y_t and trims $\mathcal{H}_t$ to $\mathcal{H}_{t+1}$ via an importance-weighted empirical risk minimization:

$$\mathcal{H}_{t+1} = \Big\{h \in \mathcal{H}_t \colon L_t(h) \leq L_t(\widehat{h}_t) + 2\Delta_t\Big\},$$

where $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$ for some fixed confidence parameter $\delta > 0$, and where $L_t(h)$ denotes the importance-weighted empirical risk of hypothesis $h \in \mathcal{H}$:

$$L_t(h) = \sum_{s=1}^{t} \frac{Q_s}{p_s} \ell(h(x_s), y_s),$$

with $\widehat{h}_t = \operatorname{argmin}_{h \in \mathcal{H}_t} L_t(h)$ its minimizer. After T rounds, IWAL returns the hypothesis $\widehat{h}_T$.

The theoretical guarantees for IWAL can be expressed in terms of the *generalized disagreement coefficient*.[1] Given $r > 0$, let $B_{\text{IWAL}}(f, r)$ denote the ball of radius r centered in $f \in \mathcal{H}$: $B_{\text{IWAL}}(f, r) = \{g \in \mathcal{H} \colon \mathcal{L}(f, g) \leq r\}$, where the distance is measured by the expected disagreement value $\mathcal{L}(f, g)$. The generalized disagreement coefficient is defined as the minimum value of θ_{IWAL} such that for all $r > 0$,

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\max_{h \in B_{\text{IWAL}}(h^*, r)} \mathcal{L}(h(x), h^*(x)) \right] \leq \theta_{\text{IWAL}} r\,. \tag{2}$$

Note that, by definition, we have $\theta_{\text{IWAL}} \geq 1$. The disagreement coefficient is a critical parameter that is widely used in the analysis of disagreement-based active learning: Hanneke (2007) proved upper and lower bounds for the label complexity of the A^2 algorithm in terms of the disagreement coefficient; Dasgupta et al. (2008) also gave an upper bound for the DHM algorithm using the disagreement coefficient. See (Hanneke, 2014) for a more extensive analysis of the disagreement coefficient and active learning algorithms.

Let $\mathcal{F}_t$ denote all the previous observations up to time t: $\mathcal{F}_t = \{(x_1, y_1, p_1, Q_1), \cdots, (x_t, y_t, p_t, Q_t)\}$ with $\mathcal{F}_0 = \emptyset$. Then, IWAL admits the following learning guarantee (Beygelzimer et al., 2009).

Theorem 1 (IWAL). *For any $\delta > 0$, with probability at least $1 - \delta$, for any $t > 0$, the following holds: (1) $h^* \in \mathcal{H}_t$; (2) for all $h \in \mathcal{H}_t$, $R(h) \leq R^* + 4\Delta_{t-1}$; (3) additionally,*

$$R(\widehat{h}_T) \leq R^* + 2\Delta_T, \tag{3}$$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \big[p_t | \mathcal{F}_{t-1}\big] \leq 4\theta_{\text{IWAL}} K_\ell \big(R^* + 2\Delta_{t-1}\big), \tag{4}$$

where $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$, and where K_ℓ is a constant that depends on the loss function ℓ:

$$K_\ell = \sup_{z, z' \in \mathcal{Z}} \left| \frac{\max_{y \in \mathcal{Y}} \ell(z, y) - \ell(z', y)}{\min_{y \in \mathcal{Y}} \ell(z, y) - \ell(z', y)} \right|.$$

Note that we have $K_\ell \geq 1$. When ℓ is the logistic loss, which is the loss used in the experiments reported by the authors of IWAL, and the output of $\mathcal{H}$ is bounded by M, that is $\mathcal{Z} \subseteq [-M, M]$, then K_ℓ can be as large as $1 + e^M$. Large values of K_ℓ in combination with $\theta_{\text{IWAL}} \geq 1$ may result in $4\theta_{\text{IWAL}} K_l R^* > 1$, which would make the label complexity bound (4) vacuous, since $p_t \leq 1$ by definition. In the next section, we introduce an improved label complexity bound for IWAL, which removes the dependency on K_l and is strictly more favorable than (4).

4. Improved Guarantees for IWAL

In this section, we present a more favorable label complexity guarantee for IWAL. To do so, we first introduce a new definition of the generalized disagreement coefficient, θ, that is a lower bound on θ_{IWAL}, and then prove a label complexity bound for IWAL in terms of θ.

For any two hypotheses $h, h' \in \mathcal{H}$, we denote by $\rho(h, h')$ their distance defined as follows in terms of their losses:

$$\rho(h, h') = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} \big[|\ell(h(x), y) - \ell(h'(x), y)|\big].$$

We denote by $B(h^*, r)$ the ball of radius $r \geq 0$: $B(h^*, r) = \{h \in \mathcal{H} \colon \rho(h, h^*) \leq r\}$. Our new disagreement coefficient is the infimum value of $\theta > 0$ such that for all $r \geq 0$:

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\max_{h \in B(h^*, r)} \mathcal{L}(h(x), h^*(x)) \right] \leq \theta r. \tag{5}$$

Although (5) is syntactically similar to (2), for the definition of θ_{IWAL}, the distance metrics defining $B(h^*, r)$ and $B_{\text{IWAL}}(h^*, r)$ are distinct: B_{IWAL} is defined in terms of a disagreement-based distance metric $\mathcal{L}(\cdot, \cdot)$, while B is defined in terms of a loss-based metric $\rho(\cdot, \cdot)$. We show that $\theta \leq \theta_{\text{IWAL}}$, and that one can derive a more favorable label complexity bound for IWAL using θ.

Theorem 2. *For any $\delta > 0$, with probability at least $1 - \delta$, for all $t \in [T]$, the following holds for the label requesting probability p_t of* IWAL*:*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \big[p_t | \mathcal{F}_{t-1}\big] \leq 4\theta \big(R^* + 2\Delta_{t-1}\big), \tag{6}$$

with $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$. Furthermore, the following inequality holds: $\theta \leq \theta_{\text{IWAL}}$.

Compared to Theorem 1, our new label complexity bound removes the dependency on K_l and replaces θ_{IWAL} with θ. The inequalities $K_l \geq 1$ and $\theta_{\text{IWAL}} \geq \theta$ show that the

[1] The generalized disagreement coefficient coincides with the standard disagreement coefficient (Hanneke, 2007) when ℓ is the zero-one loss.

expression (6) is strictly more favorable than (4). Note, K_l can in fact be very large for some losses or hypothesis sets. The proof of Theorem 2 is mostly similar to that of IWAL and is given in Appendix A.

5. Enhanced IWAL Using the Disagreement Graph

In this section, we present an enhanced version of the IWAL algorithm, IWAL-D, that exploits the disagreement graph. IWAL-D adopts the same label requesting policy as IWAL, but it prunes the hypothesis set more aggressively using disagreement-graph-based slack terms. We provide a theoretical analysis of IWAL-D in terms of the property of the disagreement graph, and show that IWAL-D admits a more favorable learning guarantee than IWAL, especially when the disagreement graph is favorable.

The motivation for IWAL-D comes from the following lemma that relates the importance-weighted empirical error and the generalization error. Due to space limitation, the proofs of all the theoretical results are given in Appendix A.

Lemma 1 (IWAL-D). *For any $\delta > 0$, with probability at least $1-\delta$, for all $t \in [T]$ and for all $f, g \in \mathcal{H}_t$, the following inequality holds:*

$$|L_t(f) - L_t(g) - R(f) + R(g)| \leq \big(1 + \mathcal{L}(f,g)\big)\Delta_t,$$

where $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$.

Lemma 1 gives a customized concentration bound for every pair (f, g) in terms of $\mathcal{L}(f, g) \leq 1$, thus, it is more refined than the corresponding concentration bound of IWAL, which admits a term $2\Delta_t$ on the right-hand side. Using Lemma 1 to prune the hypothesis sets, we obtain the pruning strategy of the IWAL-D algorithm: $\mathcal{H}_t$ is trimmed to $\mathcal{H}_{t+1}$ according to the following

$$\mathcal{H}_{t+1} = \Big\{h \in \mathcal{H}_t \colon L_t(h) \leq L_t(\widehat{h}_t) + \big(1 + \mathcal{L}(h, \widehat{h}_t)\big)\Delta_t\Big\}.$$

The full pseudocode of IWAL-D is given in Algorithm 1, with slack term $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$. Lemma 1 guarantees that, with high probability, IWAL-D always maintains the best-in-class predictor in $\mathcal{H}_t$ for all $t \in [T]$, which is necessary for the active learning algorithm to succeed. On the other hand, since IWAL-D uses a smaller slack term than IWAL for pruning ($(1 + \mathcal{L}(h, \widehat{h}_t))\Delta_t \leq 2\Delta_t$), it shrinks the hypothesis set more aggressively and therefore reduces the label complexity.

We now present the learning guarantee for IWAL-D, which is based on Lemma 1. The proof is mostly the same as the proof of IWAL except using Lemma 1 as the concentration lemma, and is thus omitted.

Theorem 3 (IWAL-D). *Let $\widehat{h}_T$ denote the hypothesis returned by* IWAL-D *after T rounds. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $t \in [T]$, the following holds: (1) $h^* \in \mathcal{H}_t$; (2) for all $h \in \mathcal{H}_t$,*

$$R(h) \leq R^* + \big(2 + \mathcal{L}(h, \widehat{h}_{t-1}) + \mathcal{L}(h, h^*)\big)\Delta_{t-1};$$

(3) additionally,

$$R(\widehat{h}_T) \leq R^* + \big(1 + \mathcal{L}(\widehat{h}_T, h^*)\big)\Delta_T, \text{ and} \tag{7}$$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}\big[p_t|\mathcal{F}_{t-1}\big] \leq 2\theta\Big[2R^* + \max_{h\in\mathcal{H}_t}\big(2 + \mathcal{L}(h, \widehat{h}_{t-1}) + \mathcal{L}(h, h^*)\big)\Delta_{t-1}\Big]. \tag{8}$$

Algorithm 1 IWAL-D($\mathcal{H}$)

$\mathcal{H}_1 \leftarrow \mathcal{H}$
for $t \in [T]$ **do**
  Receive x_t
  $p_t \leftarrow \max_{h,h' \in \mathcal{H}_t} \mathcal{L}(h(x_t), h'(x_t))$
  $Q_t \leftarrow$ BERNOULLI(p_t)
  if $Q_t = 1$ **then**
    $y_t \leftarrow$ LABEL(x_t)
    $\widehat{h}_t \leftarrow \operatorname{argmin}_{h\in\mathcal{H}_t} L_t(h)$
    $\mathcal{H}_{t+1} \leftarrow \big\{h \in \mathcal{H}_t \colon L_t(h) \leq L_t(\widehat{h}_t) + \big(1 + \mathcal{L}(h, \widehat{h}_t)\big)\Delta_t\big\}$
  end if
end for
return $\widehat{h}_T$

Comparing Theorem 3 to the guarantees of IWAL (Theorem 1), both the generalization bound (7) of $\widehat{h}_T$ and the label complexity (8) per round are improved, since $\mathcal{L}(\cdot,\cdot) \leq 1$ by definition. The improvement in the learning guarantees heavily depends on the property of the average disagreement values near h^*. When the pair of hypotheses considered, in particular the pair made of the best-in-class h^* and the empirical risk minimizer $\widehat{h}_T$ agree almost everywhere, then the generalization bound of $R(\widehat{h}_T)$ can be reduced from $R^* + 2\Delta_T$ (IWAL) to approximately $R^* + \Delta_T$ (IWAL-D). Similarly, when the average disagreement values between all pairs in $\mathcal{H}_t$ are small, the label complexity bound is smaller than that of IWAL.

However, the result of Theorem 3 is not so straightforward to digest, as it involves disagreement values which are not as clear as, for example the differences in the generalization error. To remove the disagreement values from the guarantees and to better analyze the improvement over IWAL, we introduce a new term, which we refer to as the *disagreement graph coefficient* η, and present guarantees for IWAL-D in terms of η and the generalization error only. The disagreement graph coefficient is the infimum value of η such that,

$$\forall r > 0, \max_{h\in B(h^*,r)} \mathcal{L}(h, h^*) \leq \eta r,$$

where $B(h^*, r)$ is the ball centered at h^* with radius r, as

defined in Section 4. Note that we have $\eta \geq 1$, since

$$\rho(h, h') = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[|\ell(h(x), y) - \ell(h'(x), y)|\right] \leq \mathcal{L}(h, h').$$

Furthermore, by definition $\eta \leq \theta$, since

$$\max_{h\in B(h^*,r)} \mathcal{L}(h, h^*) \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_X} \left[\max_{h\in B(h^*,r)} \mathcal{L}(h(x), h^*(x))\right],$$

and thus existing upper bounds of the disagreement coefficient under various scenarios also apply to η. In particular, when ℓ is the 0-1 loss, it is easy to show that $\eta = 1$ under all distributions. For other loss functions, for η to be small, one favorable learning scenario is where the best-in-class h^* is surrounded by an isolated cluster of hypotheses, where every hypothesis within the cluster agrees with h^* almost everywhere, thus $\mathcal{L}(h, h^*) \approx \rho(h, h^*)$ and η is close to 1.

We now derive the following learning guarantees for IWAL-D only in terms of η and the best-in-class error R^*.

Corollary 1. *Let $\widehat{h}_T$ denote the hypothesis returned by* IWAL-D *after T rounds. Then, for all $\delta > 0$, with probability at least $1 - \delta$, when $2\eta(R^* + \Delta_T) < 1$,*

$$R(\widehat{h}_T) \leq R^* + \left(\frac{1 + 2\eta R^*}{1 - \eta\Delta_T}\right)\Delta_T \leq R^* + 2\Delta_T.$$

Moreover, with probability at least $1 - \delta$, for t such that $3\eta(R^* + 2\Delta_{t-1}) < 1$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X} \left[p_t|\mathcal{F}_{t-1}\right]$$
$$\leq 4\theta\left[R^* + \left(\frac{1 + 3\eta R^*}{1 - 3\eta\Delta_{t-1}}\right)\Delta_{t-1}\right] \leq 4\theta(R^* + 2\Delta_{t-1}).$$

Corollary 1 states that, when ηR^* is small enough to meet the desired assumptions, IWAL-D admits improved guarantees over IWAL in terms of both the generalization error and label complexity. Both can be quantified in terms of the disagreement graph coefficient η and the best-in-class error R^*. The smaller ηR^* is, the larger is the improvement of IWAL-D's guarantee over that of IWAL.

It is worth emphasizing again that IWAL-D always admits more favorable learning guarantees than IWAL. When the preconditions in Corollary 1 do not hold, the improvement may not be captured by our disagreement graph coefficient analysis, but IWAL-D is still at least as favorable as IWAL.

6. Enhanced IWAL-D Using a Zooming-in Technique

In this section, we exploit another property of the average disagreements and describe an algorithm, IZOOM that further improves upon IWAL-D by adaptively enriching the hypothesis set near the current best predictor. We show that IZOOM provides an additional improvement over IWAL-D both in terms of generalization bound and label complexity.

For the IZOOM algorithm, we need additional assumptions on the loss function: (1) The loss function takes the form $\ell(z, y) = f(zy)$, for some function f; (2) Function f is non-increasing, convex, and 1-Lipschitz. Many binary classification loss functions meet these assumptions, including the logistic loss $\ell(z, y) = \log(1 + e^{-zy})$ and the hinge loss $\ell(z, y) = \max(0, 1 - yz)$. In this section, we assume that these properties hold for the loss function considered.

Recall that the IWAL-D algorithm works on a finite set of hypotheses $\mathcal{H}$. Although IWAL-D returns a final predictor that is close to the best-in-class hypothesis among $\mathcal{H}$, the best-in-class error, or equivalently the approximation error of $\mathcal{H}$ itself, may potentially be quite large. To reduce the approximation error, the learner can increase the size of $\mathcal{H}$ and, without any prior knowledge of the data distribution, sample uniformly with a finer granularity from the function space to construct $\mathcal{H}$.

Now, with access to the average disagreements between arbitrary hypotheses, which can be accurately estimated from large amounts of unlabeled data, the first attempt the learner could make is to construct $\mathcal{H}$ in a distribution-dependent way as follows: first, create an ϵ-cover $\mathcal{G}$ over the infinite function space $\mathcal{H}_\infty$, where $\mathcal{G}$ contains a set of hypotheses such that for every h in $\mathcal{H}_\infty$, there exists a $g \in \mathcal{G}$ with $\mathcal{L}(h, g) \leq \epsilon$, and thus $|R(h) - R(g)| \leq \mathcal{L}(h, g) \leq \epsilon$; next, set $\mathcal{H} = \mathcal{G}$, which gives a finite hypothesis set with the desired resolution of ϵ in the generalization error. The learner can then apply any active learning algorithm to $\mathcal{H}$ to achieve favorable learning guarantees. Depending on the data distribution, the cardinality of $\mathcal{H}$ can be significantly smaller. In Appendix B, we present the formal definition of the ϵ-cover and an example of applying IWAL to the ϵ-cover, and prove its learning guarantees in terms of ϵ.

However, with or without access to the average disagreements, the size of $\mathcal{H}$ in general increases exponentially in the dimension of the feature space, making it impractical to run any active learning algorithm with a hypothesis set $\mathcal{H}$ of a desired resolution, especially for datasets with high-dimensional feature vectors. Furthermore, it is unclear how to efficiently construct an ϵ-cover for the function space, using the average disagreements as the distance metric. Can we adaptively increase the resolution of the hypothesis set while actively requesting the labels? More importantly, can we still prove theoretical guarantees for the final output while changing the hypothesis set? In this section, we describe the IZOOM algorithm that achieves this goal.

The pseudocode of the IZOOM algorithm is given in Algorithm 2. IZOOM uses label requesting and hypothesis pruning policies similar to those of IWAL-D, with a slightly

Algorithm 2 IZOOM($\mathcal{H}$)

$\mathcal{H}_1 \leftarrow \mathcal{H}$
for $t \in [T]$ **do**
  Receive x_t
  $p_t \leftarrow \max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x_t), g(x_t)) + \frac{4}{T}$
  $Q_t \leftarrow$ BERNOULLI(p_t)
  if $Q_t = 1$ **then**
    $y_t \leftarrow$ LABEL(x_t)
    $\widehat{h}_t \leftarrow \operatorname{argmin}_{h \in \mathcal{H}_t} L_t(h)$
    $\mathcal{H}'_{t+1} \leftarrow \{h \in \mathcal{H}_t : L_t(h) \le L_t(\widehat{h}_t) + (1 + \mathcal{L}(h, \widehat{h}_t) + \frac{4}{T})\Delta_t + \frac{4}{T}\}$
    $\mathcal{H}''_{t+1} \leftarrow$ RESAMPLE($\mathcal{H}'_{t+1}, |\mathcal{H}| - |\mathcal{H}'_{t+1}|$)
    $\mathcal{H}_{t+1} \leftarrow \mathcal{H}'_{t+1} \cup \mathcal{H}''_{t+1}$
  end if
end for
return $\widehat{h}_T$

different slack term $\Delta_t = \sqrt{(2/t)\log(4T\mathcal{N}^2_{\infty,\frac{1}{T}}/\delta)}$, where $\mathcal{N}_{\infty,\epsilon}$ is the ϵ-covering number of conv($\mathcal{H}$) with respect to the ℓ_∞ norm. However, after IWAL-D has pruned the hypothesis set from $\mathcal{H}_t$ to $\mathcal{H}'_{t+1}$, IZOOM further samples new hypotheses within the convex hull of $\mathcal{H}'_{t+1}$ and combines them with $\mathcal{H}'_{t+1}$ to define $\mathcal{H}_{t+1}$. The subroutine for sampling new hypotheses is given in Algorithm 3 and illustrated in Figure 1. In this illustration, at the end of round t, IZOOM locates the empirical best predictor $\widehat{h}_t$ among $\mathcal{H}_t$, and prunes out three hypotheses that are far away from $\widehat{h}_t$ to obtain $\mathcal{H}'_{t+1}$. Next, IZOOM randomly samples three new hypotheses from conv($\mathcal{H}'_{t+1}$) by taking convex combinations of $\widehat{h}_t$ and other hypotheses in $\mathcal{H}'_{t+1}$, and combine them with $\mathcal{H}'_{t+1}$ to obtain $\mathcal{H}_{t+1}$. As IZOOM keeps *zooming-in*, thereby enriching the current hypothesis set $\mathcal{H}'_{t+1}$, we expect the final hypothesis set $\mathcal{H}_T$, from which we learn the final hypothesis output $\widehat{h}_T$, to admit a substantially smaller approximation error.

In many cases of interest, the covering number $\mathcal{N}_{\infty,\epsilon}$ in the expression of Δ_t is polynomial in $\frac{1}{\epsilon}$. In particular, when $\mathcal{H}$ is the family of functions with bounded norm in the reproducing kernel Hilbert space (RKHS) associated with Gaussian kernels, as shown by Guo et al. (1999) [Eq. (28)], we have $\log \mathcal{N}_{\infty,\epsilon} = O(\log^{\frac{3}{2}} \frac{1}{\epsilon})$. To fix ideas, we set $\epsilon = \frac{1}{T}$, but note that ϵ can be chosen based on the covering number.

By design, IZOOM maintains a fixed number of hypotheses for all $\mathcal{H}_t$, $t \in [T]$, so that IZOOM does not require additional computational resources or storage space than the original IWAL-D using the same number of hypotheses. With $|\mathcal{H}_t|$ being fixed, the more hypotheses IWAL-D prunes out from $\mathcal{H}_t$, the more hypotheses IZOOM samples and adds into $\mathcal{H}_{t+1}$. Thus, the greedy pruning strategy of IWAL-D helps IZOOM prune out and add in more hypotheses.

A key step in providing guarantees for the IZOOM algorithm is the concentration bound of Lemma 1. In particular, one needs to show that for all $t \in [T]$ and for all pairs of hypotheses $f_t, g_t \in \mathcal{H}_t$, the random variable $|\ell(f_t(x_s), y_s) - \ell(g_t(x_s), y_s)|Q_s/p_s$ is bounded for all past observations $\{(x_s, y_s, p_s, Q_s) : s \le t\}$. This naturally holds for IWAL-D since the hypothesis set $\mathcal{H}_t$ is always shrinking. However, it is much less straightforward to show that for IZOOM, since it augments $\mathcal{H}_t$ with new hypotheses. To prove the guarantees for IZOOM, we rely on a novel property of the disagreement value. Its proof uses the non-increasing property of the loss function.

Algorithm 3 RESAMPLE($\mathcal{H}, n$)

$\widehat{h}_t \leftarrow \operatorname{argmin}_{\mathcal{H}} L_t(h)$
Sample $\lambda_i \sim U[0,1], \forall i \in [n]$
Sample $h_i \sim \mathcal{H} \setminus \widehat{h}_t, \forall i \in [n]$
$\widetilde{h}_i \leftarrow \lambda_i \widehat{h}_t + (1 - \lambda_i) h_i, \forall i \in [n]$
return $\{\widetilde{h}_i : i \in [n]\}$

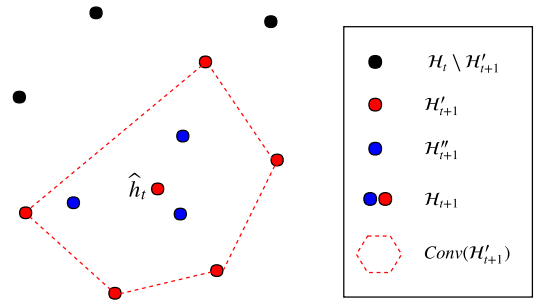

Figure 1. Illustration of IZOOM at time t. $\mathcal{H}'_{t+1}$ (the red points) is obtained by pruning out three hypotheses (the black points) from $\mathcal{H}_t$. IZOOM then samples three new hypotheses (the blue points) from conv($\mathcal{H}'_{t+1}$), and combines them with $\mathcal{H}'_{t+1}$ to get $\mathcal{H}_{t+1}$.

Lemma 2. *Let $\mathcal{H}_s$ be a set of hypotheses at time s. For any $t \ge s$, assume that $\mathcal{H}_t$ is in the convex hull of $\mathcal{H}_s$:*

$$\mathcal{H}_t \subseteq conv(\mathcal{H}_s) = \left\{ \sum_{i=1}^{|\mathcal{H}_s|} \lambda_i h_i(x) : \lambda \in \Delta, h_i \in \mathcal{H}_s \right\},$$

where Δ is the probability simplex of dimension $|\mathcal{H}_s|$. Then, for any $x \in \mathcal{X}$, the following inequality holds:

$$\max_{h,h' \in \mathcal{H}_t} \mathcal{L}(h(x), h'(x)) \le \max_{h,h' \in \mathcal{H}_s} \mathcal{L}(h(x), h'(x)). \quad (9)$$

Lemma 2 states that for the sequence of hypothesis sets $\mathcal{H}_t$ defined by IZOOM, the maximum disagreement value on a fixed point x is non-increasing despite the fact that $\mathcal{H}_t$ is augmented with newly sampled hypotheses. This property is the key to proving the desired concentration result of IZOOM, Lemma 3, from which we derive the generalization bounds and the label complexities.

Lemma 3 (IZOOM)**.** *For all $\delta > 0$, with probability at least*

$1-\delta$, *for all* $t \le T$, *and for all* $f, g \in \mathcal{H}_t$,

$$|L_t(f) - L_t(g) - R(f) + R(g)| \le \big(1 + \mathcal{L}(f,g) + \tfrac{4}{T}\big)\Delta_t + \tfrac{4}{T}.$$

Before presenting the learning guarantees of IZOOM, we introduce some additional notation. Let $\widetilde{\mathcal{H}}_t$ denote the set of hypotheses that have been considered up to round t, which includes the original $\mathcal{H}$ and the newly sampled hypotheses: $\widetilde{\mathcal{H}}_t = \mathcal{H} \cup \{\cup_{s=1}^{t} \mathcal{H}''_s\}$. Clearly $\widetilde{\mathcal{H}}_1 \subseteq \widetilde{\mathcal{H}}_2 \subseteq \cdots$, as IZOOM considers increasingly richer hypothesis sets. On the other hand, the sampling subroutine (Algorithm 3) creates random combinations of current empirical best predictor $\widehat{h}_t$ and other hypotheses in $\mathcal{H}'_{t+1}$, thus $\mathcal{H}''_{t+1} \subseteq \text{conv}(\mathcal{H}'_{t+1}) \subseteq \text{conv}(\mathcal{H}_t)$. It follows that $\mathcal{H}_{t+1} \subseteq \text{conv}(\mathcal{H}_t)$ for all t, and therefore $\mathcal{H}_t \subseteq \text{conv}(\mathcal{H}_s)$ for all $s \le t$. Let h^*_t be the best-in-class hypothesis of $\widetilde{\mathcal{H}}_t$, and let $R^*_t = R(h^*_t)$.

We now present the learning guarantees for IZOOM. Like IWAL and IWAL-D, IZOOM's label complexity bound also depends on the disagreement coefficient. However, for IZOOM, the disagreement coefficient changes per round since it is defined with respect to the changing $\mathcal{H}_t$ and h^*_t. We denote by θ_t the disagreement coefficient for round t, and give guarantees in terms of θ_t. The proof uses Lemma 3 and the convexity of the loss function.

Theorem 4 (IZOOM). *For any* $\delta > 0$, *with probability at least* $1-\delta$, *for all* $t \in [T]$, *the following holds: (1)* $h^*_t \in \mathcal{H}_t$; *(2) For all* $h \in \mathcal{H}_{t+1}$,

$$R(h) \le R^*_t + \big(2 + \mathcal{L}(h, \widehat{h}_t) + \mathcal{L}(h, h^*_t)\big)\Delta_t + \tfrac{8}{T}(1 + \Delta_t);$$

(3) Additionally,

$$R(\widehat{h}_T) \le R^*_T + \big(1 + \mathcal{L}(\widehat{h}_T, h^*_T)\big)\Delta_T + \tfrac{4}{T}(1 + \Delta_T), \text{ and}$$
$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_X}\big[p_{t+1} | \mathcal{F}_t\big] \le 2\theta_t\big[2R^*_t + \max_{h \in \mathcal{H}_{t+1}} \big(2 + \mathcal{L}(h, \widehat{h}_t) + \mathcal{L}(h, h^*_t)\big)\Delta_t\big] + \tfrac{4}{T}(1 + 4\theta_t + 4\theta_t\Delta_t).$$

Theorem 4 shows that, remarkably, the best-in-class predictor among the cumulative hypothesis set $\widetilde{\mathcal{H}}_t$ is always contained in $\mathcal{H}_t$, which is only a subset of $\widetilde{\mathcal{H}}_t$. As IZOOM enriches $\widetilde{\mathcal{H}}_t$ over time by sampling more hypotheses near $\widehat{h}_t$ (thus near h^*_t as well), the best-in-class error of $\widetilde{\mathcal{H}}_t$ will be decreasing. It follows that $R^*_T \le R^*$, where R^* is the best-in-class error of IWAL-D using the initial $\mathcal{H}$. In addition, it is reasonable to assume that for smooth distributions, the disagreement coefficient θ_t is stable and does not change dramatically over the changing $\mathcal{H}_t$. Therefore, IZOOM also decreases the label complexity over IWAL-D since the label complexity scales with the decreasing best-in-class error $R(h^*_t)$. Overall, IZOOM simultaneously decreases the generalization bound and the label complexity over IWAL-D.

To better understand the result of Theorem 4, we derive similar results to Corollary 1 for IZOOM. In the same way as for θ_t, we denote η_t the time-varying disagreement graph coefficient defined with respect to $\mathcal{H}_t$ and h^*_t. The rest of the proof is syntactically the same as the proof of Corollary 1.

Table 1. Binary classification dataset summary: number of observations (N), number of features (d), proportion of minority class (r). Datasets are ordered by number of features. For high-dimensional datasets we only use the first 10 principal components.

Dataset	N	d	r
skin	245,057	3	0.208
codrna	59,535	8	0.333
shuttle	43,500	9	0.216
magic04	19,020	10	0.352
ijcnn1	49,990	22	0.097
covtype	581,012	54	0.488
nomao	34,465	118	0.286
a9a	48,842	123	0.239

Corollary 2 (IZOOM). *For all* $\delta > 0$, *with probability at least* $1-\delta$, *when* $2\eta_T(R^*_T + \Delta_T + \frac{2}{T}(1+\Delta_T)) \le 1$,

$$R(\widehat{h}_T) \le R^*_T + \Big(\frac{1 + 2\eta_T R^*_T + \frac{4}{T}(1+\eta_T)}{1 - \eta_T\Delta_T}\Big)\Delta_T + \tfrac{4}{T} \le R^*_T + 2\Delta_T + \tfrac{4}{T}(1+\Delta_T).$$

Moreover, with probability at least $1-\delta$, *for* t *such that* $3\eta_t(R^*_t + 2\Delta_t + \frac{4}{T}(1+\Delta_t)) \le 1$,

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_X}\big[p_{t+1} | \mathcal{F}_t\big] \le 4\theta_t\Big[R^*_t + \Big(\frac{1 + 3\eta_t R^*_t + \frac{4}{T}(1+3\eta_t)}{1 - 3\eta_t\Delta_t}\Big)\Delta_t\Big] + \tfrac{4}{T}(1+4\theta_t) \le 4\theta_t(R^*_t + 2\Delta_t) + \tfrac{4}{T}(1 + 4\theta_t + 4\theta_t\Delta_t).$$

In the same way as for θ_t, it is reasonable to assume that, for smooth distributions, the coefficient η_t is stable over the changing $\mathcal{H}_t$. On the other hand, R^*_T can be significantly reduced by zooming in, thus $\eta_T R^*_T$ is likely to be substantially smaller than ηR^*. Thus, IZOOM is likely to achieve further improvements over IWAL-ZOOM, a variant that uses IWAL as the underlying subroutine when zooming in.

7. Experiments

In this section, we report the results of several experiments comparing IWAL, IWAL-D and IZOOM. We experimented with these algorithms in 8 binary classification datasets from the UCI repository. Table 1 summarizes the relevant statistics for these datasets. Due to space limitations, we only show the results for 4 datasets, and give the results for the remaining datasets in Appendix C. For high-dimensional datasets, we only kept the first 10 principal components of the original features. We used the standard logistic loss function, which is defined for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and hypothesis h by $\log(1 + e^{-yh(x)})$, which we then rescaled to $[0, 1]$.

For all algorithms, we randomly drew 3,000 hyperplanes with bounded norms as our base hypothesis set $\mathcal{H}$. In addi-

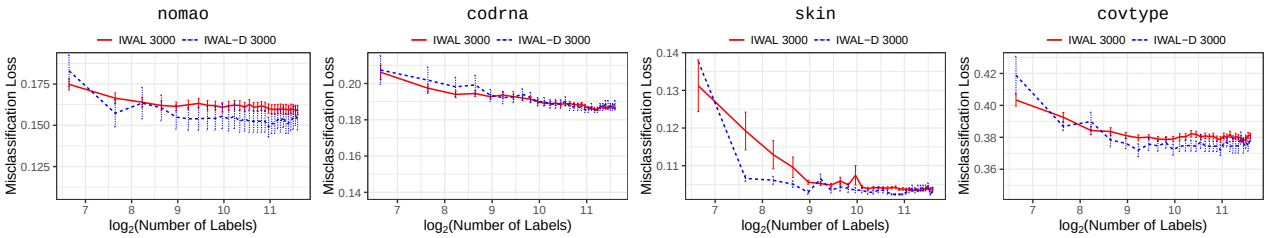

Figure 2. Misclassification loss of IWAL and IWAL-D on held-out test data versus number of labels requested ($\log_2$ scale). In some cases, we observe that IWAL-D achieves a better test performance (`nomao`, `skin`, `covtype`), sometimes mostly at the beginning. In others, the difference is not significant.

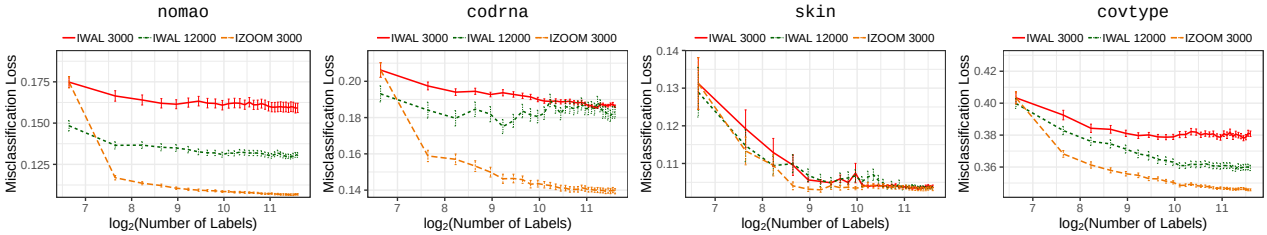

Figure 3. Misclassification loss of IWAL, and IZOOM, on held-out test data versus number of labels requested ($\log_2$ scale). The figures show that IZOOM provides significant improvements in performance over IWAL, despite it is run with just $|\mathcal{H}| = 3000$. The red curves for IWAL with $|\mathcal{H}| = 3000$ are repetitions from Figure 2.

tion, for the IZOOM algorithm we kept $|\mathcal{H}| = 3{,}000$ throughout the learning. We found that in the end, the total number of hypotheses ever considered by IZOOM was around $6{,}000$, thus for fair comparisons, we also experimented with IWAL using a larger set of $|\mathcal{H}| = 12{,}000$ hypotheses (twice as many as the number of hypotheses considered by IZOOM), since starting off with a randomly sampled hypothesis set, with no adaptation, makes IWAL intuitively less effective. For every pair of $h, h' \in \mathcal{H}$, we approximated $\mathcal{L}(h, h')$ with the average disagreement values on $2{,}000$ unlabeled samples. For each experiment, we randomly shuffled the dataset and ran the algorithms on the first 50% of the data, and tested the learned classifier on the remaining 50% of the data. We repeated the process 50 times for each dataset, and reported the average results with standard errors.

We first compared IWAL-D with IWAL, where both algorithms have $3{,}000$ initial hypotheses. Figure 2 shows the misclassification loss of IWAL and IWAL-D on held-out test data against the number of labels requested (on $\log_2$ scale). We observe that, in some cases IWAL-D outperforms IWAL with a lower misclassification error (`nomao`, `skin`, `covtype`). In other cases, the difference is not significant. Overall, there appears to be advantages in using the more aggressive disagreement-graph-based pruning strategy.

Next, we compared IZOOM to IWAL for various hypothesis set sizes. Figure 3 plots the misclassification loss on held-out test data against the number of labels (on a $\log_2$ scale). On almost all datasets, the performance of IWAL improves significantly from $|\mathcal{H}| = 3{,}000$ to $|\mathcal{H}| = 12{,}000$, as expected. However, the IZOOM algorithm with $|\mathcal{H}| = 3{,}000$ achieves almost from the beginning a significantly better prediction accuracy than the original IWAL even with the large size of $|\mathcal{H}|$ on almost all datasets. Meanwhile, IZOOM had considered a total of $6{,}000$ hypotheses within each experiment, which is only half of the largest size of $|\mathcal{H}|$ for IWAL. This again illustrates that IZOOM samples from the function space more effectively than uniformly sampling. On several datasets, the learning curve of IZOOM is much steeper than IWAL at the beginning, which makes IZOOM a promising active learning algorithm, since the performance in the early regime is of particular interest in active learning.

8. Conclusion

We presented two active learning algorithms exploiting average disagreements between hypotheses. We showed that they benefit from favorable generalization and label complexity guarantees. We also reported the results of several experiments demonstrating that they can achieve substantial performance improvements over existing active learning algorithms such as IWAL. Altogether, our theory, algorithms, and empirical results provide a very effective solution for active learning.

Acknowledgments

We thank all anonymous reviewers for their comments. This work was partly funded by NSF CCF-1535987 and NSF IIS-1618662.

References

P. Awasthi, M.-F. Balcan, and P. Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458. ACM, 2014.

P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190, 2015.

M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

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

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, pages 199–207, 2010.

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

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Region-based active learning. In *AISTATS 2019*, 2019.

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *International Conference on Computational Learning Theory*, pages 249–263. Springer, 2005.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pages 353–360, 2008.

Y. Guo, P. L. Bartlett, J. Shawe-Taylor, and R. C. Williamson. Covering numbers for support vector machines. In *COLT*, 1999.

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

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

C. Zhang. Efficient active learning of sparse halfspaces. In *Conference on Learning Theory*, 2018.

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

A. Proof

Theorem 2. *For any $\delta > 0$, with probability at least $1-\delta$, for all $t \in [T]$, the following holds for the label requesting probability p_t of* IWAL*:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[p_t|\mathcal{F}_{t-1}\right] \leq 4\theta\big(R^* + 2\Delta_{t-1}\big),$$

with $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$. Furthermore, the following inequality holds: $\theta \leq \theta_{\text{IWAL}}$.

Proof. For all $h \in \mathcal{H}$,

$$\begin{aligned}\rho(h,h^*) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[|\ell(h(x),y) - \ell(h^*(x),y)|\right] \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(h(x),y) + \ell(h^*(x),y)\right] \\ &= R(h) + R(h^*).\end{aligned}$$

From the proof of IWAL, for all $t > 0$, $h^* \in \mathcal{H}_t$, and $R(h) \leq R(h^*) + 4\Delta_{t-1}$ for all $h \in \mathcal{H}_t$. Thus, $\forall h \in \mathcal{H}_t$,

$$\rho(h,h^*) \leq R(h) + R(h^*) \leq 2R(h^*) + 4\Delta_{t-1}.$$

We can then choose $r_t = 2R^* + 4\Delta_{t-1}$ so that $\mathcal{H}_t \subseteq B(h^*, r_t)$. Then, for all $f, g \in \mathcal{H}_t$,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[p_t|\mathcal{F}_{t-1}] &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{f,g\in\mathcal{H}_t}\mathcal{L}(f(x),g(x))\right] \\ &\leq 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{h\in\mathcal{H}_t}\mathcal{L}(h(x),h^*(x))\right] \\ &\leq 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{h\in B(h^*,r_t)}\mathcal{L}(h(x),h^*(x))\right] \\ &\leq 2\theta r_t.\end{aligned}$$

Plugging in the value of r_t gives (6). Next, we show that $\theta \leq \theta_{\text{IWAL}}$. For all pairs h, h',

$$\begin{aligned}&\rho(h,h') \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[|\ell(h(x),y) - \ell(h'(x),y)|\right] \\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{y\in\mathcal{Y}}|\ell(h(x),y) - \ell(h'(x),y)|\right] = \mathcal{L}(h,h').\end{aligned}$$

Thus, $B(h^*,r) \subseteq B_{\text{IWAL}}(h^*,r)$. It follows that for any $r > 0$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{h\in B(h^*,r)}\mathcal{L}(h(x),h^*(x))\right] \\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{h\in B_{\text{IWAL}}(h^*,r)}\mathcal{L}(h(x),h^*(x))\right] \leq \theta_{\text{IWAL}}\, r.\end{aligned}$$

Since θ is the minimum value such that for all $r > 0$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\max_{h\in B(h^*,r)}\mathcal{L}(h(x),h^*(x))\right] \leq \theta r.$$

Thus, by definition $\theta \leq \theta_{\text{IWAL}}$. □

Lemma 1. *For any $\delta > 0$, with probability at least $1-\delta$, for all $t \in [T]$ and for all $f, g \in \mathcal{H}_t$, the following inequality holds:*

$$|L_t(f) - L_t(g) - R(f) + R(g)| \leq \big(1 + \mathcal{L}(f,g)\big)\Delta_t,$$

where $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$.

Proof. Fix time t and a pair of hypotheses $f, g \in \mathcal{H}_t$. For any $s \leq t$, let

$$Z_s = \frac{Q_s}{p_s}[\ell(f(x_s),y_s) - \ell(g(x_s),y_s)] - [R(f) - R(g)].$$

Since $\mathcal{H}_t$ keeps shrinking, that is, $\mathcal{H}_t \subseteq \mathcal{H}_{t-1} \cdots \subseteq \mathcal{H}_1$, by the definition of p_s, for all $s \leq t$,

$$\max_{f,g\in\mathcal{H}_t}\mathcal{L}(f(x_s),g(x_s)) \leq \max_{f,g\in\mathcal{H}_s}\mathcal{L}(f(x_s),g(x_s)) = p_s.$$

In addition,

$$\begin{aligned}|R(f) - R(g)| &= \left|\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\ell(f(x),y) - \ell(g(x),y)\right]\right| \\ &\leq \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left|\ell(f(x),y) - \ell(g(x),y)\right| \\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\mathcal{L}(f(x),g(x))\right] \\ &= \mathcal{L}(f,g)\,.\end{aligned}$$

It follows that, for the fixed pair $f, g \in \mathcal{H}_t$,

$$|Z_s| \leq \frac{\mathcal{L}(f(x_s),g(x_s))}{p_s} + |R(f) - R(g)| \leq 1 + \mathcal{L}(f,g).$$

Furthermore,

$$\begin{aligned}&\mathop{\mathbb{E}}_{\substack{Q_s\sim\text{Bernoulli}(p_s)\\(x_s,y_s)\sim\mathcal{D}}}[Z_s|\mathcal{F}_{s-1}] \\ &= \mathop{\mathbb{E}}_{(x_s,y_s)\sim\mathcal{D}}[\ell(f(x_s),y_s) - \ell(g(x_s),y_s] - (R(f) - R(g)) \\ &= 0.\end{aligned}$$

Thus, $Z_1, \cdots, Z_t$ is a bounded martingale difference sequence. Applying Azuma's inequality to $\sum_{s=1}^t Z_s/\big(1 + \mathcal{L}(f,g)\big)$,

$$\begin{aligned}&\mathbb{P}\left(|L_t(f) - L_t(g) - R(f) + R(g)| \geq \big(1 + \mathcal{L}(f,g)\big)\Delta_t\right) \\ &\quad= \mathbb{P}\left(\frac{1}{t}\left|\sum_{s=1}^t Z_s/\big(1 + \mathcal{L}(f,g)\big)\right| \geq \Delta_t\right) \\ &\quad\leq 2e^{\frac{-t\Delta_t^2}{2}} = \frac{\delta}{t(t+1)|\mathcal{H}|^2}.\end{aligned}$$

A union bound over all $t \geq 0$ and all pairs $f, g \in \mathcal{H}$ concludes the proof. □

Corollary 1. *Let $\widehat{h}_T$ denote the hypothesis returned by* IWAL-D *after T rounds. Then, for all $\delta > 0$, with probability at least $1-\delta$, when $2\eta(R^* + \Delta_T) < 1$,*

$$R(\widehat{h}_T) \le R^* + \Big(\frac{1+2\eta R^*}{1-\eta\Delta_T}\Big)\Delta_T \le R^* + 2\Delta_T.$$

Moreover, with probability at least $1-\delta$, for t such that $3\eta(R^* + 2\Delta_{t-1}) < 1$,

$$\mathbb{E}_{x\sim\mathcal{D}_X}\big[p_t|\mathcal{F}_{t-1}\big] \le 4\theta\Big[R^* + \Big(\frac{1+3\eta R^*}{1-3\eta\Delta_{t-1}}\Big)\Delta_{t-1}\Big] \le 4\theta(R^* + 2\Delta_{t-1}).$$

Proof. Since $\rho(\widehat{h}_T, h^*) \le R(h^*) + R(\widehat{h}_T)$, from the definition of η we have

$$\mathcal{L}(\widehat{h}_T, h^*) \le \eta\rho(\widehat{h}_T, h^*) \le \eta\big(2R^* + (1 + \mathcal{L}(\widehat{h}_T, h^*))\Delta_T\big).$$

Rearranging terms, when T is large enough such that $2\eta(R^* + \Delta_T) \le 1$, we get

$$\mathcal{L}(\widehat{h}_T, h^*) \le \frac{2\eta R^* + \eta\Delta_T}{1-\eta\Delta_T} \le 1\,.$$

Plugging the upper bound of $\mathcal{L}(\widehat{h}_T, h^*)$ into (7) allows us to conclude that

$$R(\widehat{h}_T) \le R^* + \Big(\frac{1+2\eta R^*}{1-\eta\Delta_T}\Big)\Delta_T \le R^* + 2\Delta_T.$$

For the label complexity bound, again by the definition of η, for all $h \in \mathcal{H}_t$:

$$\begin{aligned}
&\max_{h\in\mathcal{H}_t}\Big(\mathcal{L}(h,\widehat{h}_{t-1}) + \mathcal{L}(h,h^*)\Big)\\
&\le \max_{h\in\mathcal{H}_t}\Big(\mathcal{L}(h,h^*) + \mathcal{L}(h^*,\widehat{h}_{t-1}) + \mathcal{L}(h,h^*)\Big)\\
&\le 3\max_{h\in\mathcal{H}_t}\mathcal{L}(h,h^*) \le 3\eta\Big(\max_{h\in\mathcal{H}_t}R(h) + R(h^*)\Big)\\
&\le 3\eta\Big(2R(h^*) + \max_{h\in\mathcal{H}_t}\big(2 + \mathcal{L}(h,\widehat{h}_{t-1}) + \mathcal{L}(h,h^*)\big)\Delta_{t-1}\Big).
\end{aligned}$$

Rearranging terms, when $3\eta(R^* + 2\Delta_{t-1}) \le 1$, we have

$$\max_{h\in\mathcal{H}_t}\Big(\mathcal{L}(h,\widehat{h}_{t-1}) + \mathcal{L}(h,h^*)\Big) \le \frac{6\eta(R^* + \Delta_{t-1})}{1-3\eta\Delta_{t-1}} \le 2.$$

Plugging into (8), we get the label complexity bound in terms of η and R^*:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}\big[p_t|\mathcal{F}_{t-1}\big] &\le 4\theta\Big[R^* + \Big(\frac{1+3\eta R^*}{1-3\eta\Delta_{t-1}}\Big)\Delta_{t-1}\Big]\\
&\le 4\theta(R^* + 2\Delta_{t-1}).
\end{aligned}$$

□

Lemma 2. *Let $\mathcal{H}_s$ be a set of hypotheses at time s. For any $t \ge s$, assume that $\mathcal{H}_t$ is in the convex hull of $\mathcal{H}_s$:*

$$\mathcal{H}_t \subseteq \mathit{conv}(\mathcal{H}_s) = \Big\{\sum_{i=1}^{|\mathcal{H}_s|}\lambda_i h_i(x)\colon \lambda\in\Delta, h_i\in\mathcal{H}_s\Big\},$$

where Δ is the probability simplex of dimension $|\mathcal{H}_s|$. Then, for any $x \in \mathcal{X}$,

$$\max_{h,h'\in\mathcal{H}_t}\mathcal{L}(h(x),h'(x)) \le \max_{h,h'\in\mathcal{H}_s}\mathcal{L}(h(x),h'(x)).$$

Proof. For any $s < t$, pick two hypotheses $\widetilde{h}_1, \widetilde{h}_2 \in \mathcal{H}_t \subseteq \mathrm{conv}(\mathcal{H}_s)$. Let $\widetilde{h}_1 = \sum_{i=1}^{|\mathcal{H}_s|}\lambda_{1i}h_i$, and $\widetilde{h}_2 = \sum_{i=1}^{|\mathcal{H}_s|}\lambda_{2i}h_i$, where $\mathcal{H}_s = \{h_1, h_2, \cdots, h_{|\mathcal{H}_s|}\}$, and $\lambda_1, \lambda_2 \in \Delta$. By assumption, $\ell(h(x), y) = f(h(x)y)$ for a non-increasing function f. Thus, for any x, y,

$$\begin{aligned}
\ell(\widetilde{h}_1(x), y) &= f\Big(\sum_{i=1}^{|\mathcal{H}_s|}\lambda_{1i}h_i(x)y\Big) \le \max_{h_i\in\mathcal{H}_s} f\big(h_i(x)y\big),\\
\ell(\widetilde{h}_2(x), y) &= f\Big(\sum_{i=1}^{|\mathcal{H}_s|}\lambda_{2i}h_i(x)y\Big) \ge \min_{h_i\in\mathcal{H}_s} f\big(h_i(x)y\big),
\end{aligned}$$

where the two inequalities follow from the non-increasing property of f. Thus, all $y \in \mathcal{Y}$ and $\widetilde{h}_1, \widetilde{h}_2 \in \mathcal{H}_t$,

$$\begin{aligned}
&\ell(\widetilde{h}_1(x), y) - \ell(\widetilde{h}_2(x), y)\\
&\le \max_{h\in\mathcal{H}_s} f\big(h(x)y\big) - \min_{h\in\mathcal{H}_s} f\big(h(x)y\big)\\
&\le \max_{h,h'\in\mathcal{H}_s} f(h(x)y) - f(h'(x)y)\\
&= \max_{h,h'\in\mathcal{H}_s} \ell(h(x),y) - \ell(h'(x),y).
\end{aligned}$$

Taking a maximum over y and $\widetilde{h}_1, \widetilde{h}_2$ concludes the proof. □

The remaining proofs of this section are given in the general case of an arbitrary ϵ and its associated ϵ-cover. But, as indicated in the main body of this paper, to fix ideas, we set $\epsilon = 1/T$ to derive our main statements.

Lemma 3. *For all $\delta > 0$, with probability at least $1-\delta$, for all $t \le T$, and for all $f, g \in \mathcal{H}_t$,*

$$|L_t(f) - L_t(g) - R(f) + R(g)| \le \big(1 + \mathcal{L}(f,g) + \tfrac{4}{T}\big)\Delta_t + \tfrac{4}{T}.$$

Proof. Let $d(h,h') = \max_{x\in\mathcal{X}}|h(x) - h'(x)|$ denote ℓ_∞ distance of h and h' on $\mathcal{X}$. Let $\mathcal{G}$ denote the minimal ϵ-cover of the convex hull of $\mathcal{H}$, so that for all $h \in \mathrm{conv}(\mathcal{H})$, there exists $h' \in \mathcal{G}$, such that $d(h,h') \le \epsilon$. Let $\mathcal{N}_{\infty,\epsilon} = |\mathcal{G}|$ denote the ϵ-covering number.

Fix a time t and a pair $f, g \in \mathcal{H}_t$. Let $f', g' \in \mathcal{G}$ denote a corresponding pair of hypotheses in the ϵ-cover, such that $d(f, f') \le \epsilon$, $d(g, g') \le \epsilon$. Since the loss function is 1-Lipschitz, we have

$$\begin{aligned}
&|L_t(f) - L_t(g) - R(f) + R(g)| \\
&\le |L_t(f') - L_t(g') - R(f') + R(g') + L_t(f) - L_t(f') + \\
&+ L_t(g') - L_t(g)| + 2\epsilon = \frac{1}{t}\Big|\sum_{s=1}^{t} Z_s\Big| + 2\epsilon, \qquad (10)
\end{aligned}$$

where

$$\begin{aligned}
Z_s = &\frac{Q_s}{p_s}\big[\ell(f'(x_s), y) - \ell(g'(x_s), y_s)\big] - [R(f') - R(g')] \\
&+ \frac{Q_s}{p_s}\big[\ell(f(x_s), y) - \ell(f'(x_s), y)\big] \\
&+ \frac{Q_s}{p_s}\big[\ell(g'(x_s), y) - \ell(g(x_s), y)\big].
\end{aligned}$$

Note that

$$\begin{aligned}
&\Big|\sum_{s=1}^{t} Z_s\Big| \\
&\le \max\left\{\Big|\sum_{s=1}^{t}\Big(Z'_s + \frac{2\epsilon Q_s}{p_s}\Big)\Big|, \Big|\sum_{s=1}^{t}\Big(Z'_s - \frac{2\epsilon Q_s}{p_s}\Big)\Big|\right\},
\end{aligned}$$

where we use the shorthand $Z'_s = \frac{Q_s}{p_s}\big[\ell(f'(x_s), y) - \ell(g'(x_s), y_s)\big] - [R(f') - R(g')]$. We will then upper bound both terms in the above maximum. First,

$$\begin{aligned}
&\Big|Z'_s + \frac{2\epsilon Q_s}{p_s} - 2\epsilon\Big| \\
&= \Big|\frac{Q_s}{p_s}[\ell(f'(x_s), y) - \ell(g'(x_s), y_s) + 2\epsilon] \\
&\qquad - R(f') + R(g') - 2\epsilon\Big| \\
&\le \frac{\max_{f,g\in\mathcal{H}_t} \mathcal{L}(f(x_s), g(x_s)) + 4\epsilon}{p_s} + \mathcal{L}(f, g) + 4\epsilon \\
&\le \frac{\max_{f,g\in\mathcal{H}_s} \mathcal{L}(f(x_s), g(x_s)) + 4\epsilon}{p_s} + \mathcal{L}(f, g) + 4\epsilon \\
&\le 1 + \mathcal{L}(f, g) + 4\epsilon,
\end{aligned}$$

where the second to last inequality follows from Lemma 2 since $\mathcal{H}_t \subseteq \mathrm{conv}(\mathcal{H}_s)$, and the last inequality follows from the definition of p_s. Furthermore, from the proof of Lemma 1, $Z'_1 + \frac{2\epsilon Q_1}{p_1} - 2\epsilon, \cdots, Z'_t + \frac{2\epsilon Q_t}{p_t} - 2\epsilon$ is a martingale difference sequence with bounded absolute value. Applying Azuma's inequality to $\sum_{s=1}^{t} \frac{Z'_s + \frac{2\epsilon Q_s}{p_s} - 2\epsilon}{1 + \mathcal{L}(f,g) + 4\epsilon}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s + \frac{2\epsilon Q_s}{p_s} - 2\epsilon\Big)\Big| \ge (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t\right) \\
&= \mathbb{P}\left(\frac{1}{t}\sum_{s=1}^{t}\Big|\frac{Z'_s + \frac{2\epsilon Q_s}{p_s} - 2\epsilon}{1 + \mathcal{L}(f, g) + 4\epsilon}\Big| \ge \Delta_t\right) \le 2e^{-\frac{t\Delta_t^2}{2}}.
\end{aligned}$$

Similarly,

$$\Big|Z'_s - \frac{2\epsilon Q_s}{p_s} + 2\epsilon\Big| \le 1 + \mathcal{L}(f, g) + 4\epsilon,$$

and

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s - \frac{2\epsilon Q_s}{p_s} + 2\epsilon\Big)\Big| \ge (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t\right) \\
&= \mathbb{P}\left(\frac{1}{t}\sum_{s=1}^{t}\Big|\frac{Z'_s - \frac{2\epsilon Q_s}{p_s} + 2\epsilon}{1 + \mathcal{L}(f, g) + 4\epsilon}\Big| \ge \Delta_t\right) \le 2e^{-\frac{t\Delta_t^2}{2}}.
\end{aligned}$$

Setting $2e^{-\frac{t\Delta_t^2}{2}} = \frac{\delta}{2}$ gives $\Delta_t = \sqrt{(2/t)\log(4/\delta)}$. Thus, with probability at least $1 - \delta$,

$$\begin{aligned}
&\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s + \frac{2\epsilon Q_s}{p_s} - 2\epsilon\Big)\Big| \le (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t, \\
&\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s - \frac{2\epsilon Q_s}{p_s} + 2\epsilon\Big)\Big| \le (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s + \frac{2\epsilon Q_s}{p_s}\Big)\Big| \le (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t + 2\epsilon, \\
&\frac{1}{t}\Big|\sum_{s=1}^{t}\Big(Z'_s - \frac{2\epsilon Q_s}{p_s}\Big)\Big| \le (1 + \mathcal{L}(f, g) + 4\epsilon)\Delta_t + 2\epsilon.
\end{aligned}$$

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

$$\frac{1}{t}\Big|\sum_{s=1}^{t} Z_s\Big| \le (1 + \mathcal{L}(f, g) + 4\epsilon)\sqrt{(2/t)\log(4/\delta)} + 2\epsilon.$$

Combining the inequality above with (10), we have that, for this fixed pair $(f, g) \in \mathcal{H}_t$,

$$\begin{aligned}
&|L_t(f) - L_t(g) - R(f) + R(g)| \\
&\le (1 + \mathcal{L}(f, g) + 4\epsilon)\sqrt{(2/t)\log(4/\delta)} + 4\epsilon. \qquad (11)
\end{aligned}$$

Next, note that each pair $(f, g) \in \mathcal{H}_t$ is mapped to (f', g') in the ϵ-cover $\mathcal{G}$, and the Azuma inequalities are applied to (f', g') rather than (f, g). Thus, for inequality (11) to hold for all possible pairs $(f, g) \in \mathcal{H}_t$, it suffices to take a union bound over all pairs $(f', g') \in \mathcal{G}$. Finally, by taking a union bound over time $t \in [T]$, we conclude that with probability at least $1 - \delta$, for all $t \in [T]$ and all pairs $(f, g) \in \mathcal{H}_t$,

$$\begin{aligned}
&|L_t(f) - L_t(g) - R(f) + R(g)| \\
&\le (1 + \mathcal{L}(f, g) + 4\epsilon)\sqrt{(2/t)\log(4T\mathcal{N}^2_{\infty,\epsilon}/\delta)} + 4\epsilon.
\end{aligned}$$

□

Theorem 4. *For any $\delta > 0$, with probability at least $1-\delta$, for all $t \in [T]$, the following holds: (1) $h_t^* \in \mathcal{H}_t$; (2) For all $h \in \mathcal{H}_{t+1}$,*

$$R(h) \le R_t^* + \big(2 + \mathcal{L}(h, \widehat{h}_t) + \mathcal{L}(h, h_t^*)\big)\Delta_t + \tfrac{8}{T}(1+\Delta_t);$$

(3) Additionally,

$$R(\widehat{h}_T) \le R_T^* + \big(1 + \mathcal{L}(\widehat{h}_T, h_T^*)\big)\Delta_T + \tfrac{4}{T}(1+\Delta_T), \text{ and}$$
$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[p_{t+1}|\mathcal{F}_t\big] \le 2\theta_t\big[2R_t^* + \max_{h\in\mathcal{H}_{t+1}}\big(2 + \mathcal{L}(h,\widehat{h}_t) + \mathcal{L}(h,h_t^*)\big)\Delta_t\big] + \tfrac{4}{T}(1 + 4\theta_t + 4\theta_t\Delta_t).$$

Proof. We first show by induction that $h_t^* \in \mathcal{H}_t$ for all $t \in [T]$. It clearly holds for $t = 1$. Now suppose it holds for time t, that is, $h_t^* \in \mathcal{H}_t$. By Lemma 3,

$$\begin{aligned}
&L_t(h_t^*) - L_t(\widehat{h}_t)\\
&\le R(h_t^*) - R(\widehat{h}_t) + (1 + \mathcal{L}(\widehat{h}_t, h_t^*) + 4\epsilon)\Delta_t + 4\epsilon\\
&\le (1 + \mathcal{L}(\widehat{h}_t, h_t^*) + 4\epsilon)\Delta_t + 4\epsilon.
\end{aligned}$$

Thus, according to the pruning rule, $h_t^* \in \mathcal{H}'_{t+1}$. Observe that h_{t+1}^* is either the same as h_t^*, thus $h_{t+1}^* \in \mathcal{H}'_{t+1}$, or $h_{t+1}^* \in \mathcal{H}''_{t+1}$. In both cases, $h_{t+1}^* \in \mathcal{H}_{t+1} = \mathcal{H}'_{t+1} \cup \mathcal{H}''_{t+1}$, which completes the induction. Thus, $h_t^* \in \mathcal{H}_t$ for all $t \le T$. It follows from Lemma 3 that,

$$\begin{aligned}
&R(\widehat{h}_T) - R(h_T^*)\\
&\le L_T(\widehat{h}_T) - L_T(h_T^*) + (1 + \mathcal{L}(\widehat{h}_T, h_T^*) + 4\epsilon)\Delta_T + 4\epsilon\\
&\le (1 + \mathcal{L}(\widehat{h}_T, h_T^*) + 4\epsilon)\Delta_T + 4\epsilon.
\end{aligned}$$

To prove the second statement, we consider $\mathcal{H}'_{t+1}$ and $\mathcal{H}''_{t+1}$ separately. By Lemma 3 again and the pruning rule of IZOOM, for all $h \in \mathcal{H}'_{t+1}$,

$$\begin{aligned}
&R(h) - R(h_t^*)\\
&\le L_t(h) - L_t(h_t^*) + (1 + \mathcal{L}(h, h_t^*) + 4\epsilon)\Delta_t + 4\epsilon\\
&\le L_t(\widehat{h}_t) + (1 + \mathcal{L}(\widehat{h}_t, h) + 4\epsilon)\Delta_t - L_t(h_t^*)\\
&\qquad + (1 + \mathcal{L}(h, h_t^*) + 4\epsilon)\Delta_t + 8\epsilon\\
&\le \big(2 + \mathcal{L}(\widehat{h}_t, h) + \mathcal{L}(h, h_t^*) + 8\epsilon\big)\Delta_t + 8\epsilon.
\end{aligned}$$

Next we consider $\mathcal{H}''_{t+1}$. Let $h = \sum_{i:\ h_i \in \mathcal{H}'_{t+1}} \lambda_i h_i$ be a hypothesis in $\mathcal{H}''_{t+1}$, where λ is in the probability simplex of dimension $|\mathcal{H}'_{t+1}|$. Then, by the convexity of the loss function,

$$R(h) \le \sum_{i:\ h_i \in \mathcal{H}'_{t+1}} \lambda_i R(h_i) \le \max_{h\in\mathcal{H}'_{t+1}} R(h)$$

Thus, for all $h \in \mathcal{H}_{t+1}$,

$$R(h) \le R(h_t^*) + (2 + \mathcal{L}(\widehat{h}_t, h) + \mathcal{L}(h, h_t^*) + 8\epsilon)\Delta_t + 8\epsilon.$$

The label complexity bound follows from the definition of the disagreement coefficient θ_t. For more details, see the proof of Theorem 2.

□

Corollary 2. *For all $\delta > 0$, with probability at least $1-\delta$, when $2\eta_T(R_T^* + \Delta_T + \frac{2}{T}(1+\Delta_T)) \le 1$ we have*

$$\begin{aligned}
R(\widehat{h}_T) &\le R_T^* + \Big(\frac{1 + 2\eta_T R_T^* + \frac{4}{T}(1+\eta_T)}{1 - \eta_T\Delta_T}\Big)\Delta_T + \tfrac{4}{T}\\
&\le R_T^* + 2\Delta_T + \tfrac{4}{T}(1+\Delta_T)\,.
\end{aligned}$$

Moreover, with probability at least $1-\delta$, for t such that $3\eta_t(R_t^* + 2\Delta_t + \frac{4}{T}(1+\Delta_t)) \le 1$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[p_{t+1}|\mathcal{F}_t\big]\\
&\le 4\theta_t\Big[R_t^* + \Big(\frac{1 + 3\eta_t R_t^* + \frac{4}{T}(1+3\eta_t)}{1 - 3\eta_t\Delta_t}\Big)\Delta_t\Big] + \tfrac{4}{T}(1+4\theta_t)\\
&\le 4\theta_t(R_t^* + 2\Delta_t) + \tfrac{4}{T}(1 + 4\theta_t + 4\theta_t\Delta_t).
\end{aligned}$$

Proof. Since $\rho(\widehat{h}_T, h_T^*) \le R(h_T^*) + R(\widehat{h}_T)$, from the definition of η_T we have

$$\begin{aligned}
\mathcal{L}(\widehat{h}_T, h_T^*) &\le \eta_T\rho(\widehat{h}_T, h_T^*)\\
&\le \eta_T\big(2R_T^* + (1 + \mathcal{L}(\widehat{h}_T, h_T^*) + 4\epsilon)\Delta_T + 4\epsilon\big).
\end{aligned}$$

When T is large enough, such that $2\eta_T(R_T^* + \Delta_T + 2\epsilon(1+\Delta_T)) \le 1$, we get

$$\mathcal{L}(\widehat{h}_T, h^*) \le \frac{\eta_T(2R_T^* + (1+4\epsilon)\Delta_T + 4\epsilon)}{1 - \eta_T\Delta_T} \le 1\,.$$

Plugging the upper bound of $\mathcal{L}(\widehat{h}_T, h^*)$ into Theorem 4 allows us to conclude that

$$\begin{aligned}
R(\widehat{h}_T) &\le R_T^* + \Big(\frac{1 + 2\eta_T R_T^* + 4\epsilon(1+\eta_T)}{1 - \eta_T\Delta_T}\Big)\Delta_T + 4\epsilon\\
&\le R_T^* + 2\Delta_T + 4\epsilon(1+\Delta_T).
\end{aligned}$$

For the label complexity bound, again by the definition of η_t, for all $h \in \mathcal{H}_{t+1}$:

$$\begin{aligned}
&\max_{h\in\mathcal{H}_{t+1}}\Big(\mathcal{L}(h,\widehat{h}_t) + \mathcal{L}(h,h_t^*)\Big)\\
&\le \max_{h\in\mathcal{H}_{t+1}}\Big(\mathcal{L}(h,h_t^*) + \mathcal{L}(h_t^*,\widehat{h}_t) + \mathcal{L}(h,h_t^*)\Big)\\
&\le 3\max_{h\in\mathcal{H}_{t+1}}\mathcal{L}(h,h_t^*) \le 3\eta_t\Big(\max_{h\in\mathcal{H}_{t+1}} R(h) + R_t^*\Big)\\
&\le 3\eta_t\Big(2R_t^*\\
&+ \max_{h\in\mathcal{H}_{t+1}}\big(2 + \mathcal{L}(h,\widehat{h}_t) + \mathcal{L}(h,h_t^*) + 8\epsilon\big)\Delta_t + 8\epsilon\Big).
\end{aligned}$$

Rearranging terms, when $3\eta_t(R_t^* + 2\Delta_t + 4\epsilon(1+\Delta_t)) \le 1$, we have

$$\max_{h\in\mathcal{H}_{t+1}} \Big(\mathcal{L}(h,\widehat{h}_t) + \mathcal{L}(h,h_t^*)\Big) \le \frac{6\eta_t(R_t^* + (1+4\epsilon)\Delta_t + 4\epsilon)}{1-3\eta_t\Delta_t} \le 2.$$

Plugging into Theorem 4, we get the label complexity bound in terms of η_t and R_t^*:

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[p_{t+1}|\mathcal{F}_t\big]\\ &\le 4\theta_t\Big[R_t^* + \Big(\frac{1+3\eta_t R_t^* + 4\epsilon(1+3\eta_t)}{1-3\eta_t\Delta_t}\Big)\Delta_t\Big] + 4\epsilon(1+4\theta_t)\\ &\le 4\theta_t(R_t^* + 2\Delta_t) + 4\epsilon(1+4\theta_t+4\theta_t\Delta_t).\end{aligned}$$

□

B. IWAL with ϵ-cover

As discussed before, the main advantage of the IZOOM algorithm is that it does not require the learner to have access to the ϵ-cover, which is computationally expensive to construct in general. However, if the learner can efficiently construct an ϵ-cover, then she can run any active learning algorithms with the desired resolution ϵ. In this section, we give a definition of the ϵ-cover with respect to the disagreement metric, and prove learning guarantees for running IWAL with such an ϵ-cover.

Definition 1. *We say $\mathcal{G}\subseteq\mathcal{H}_\infty$ is an ϵ-cover of a infinite hypothesis set $\mathcal{H}_\infty$ with respect to the disagreement metric $\mathcal{L}(\cdot,\cdot)$, if for all $h\in\mathcal{H}_\infty$, there exists $g\in\mathcal{G}$ with $\mathcal{L}(h,g)\le\epsilon$. We define by $N(\mathcal{H}_\infty,\epsilon)$ the cardinality of the smallest ϵ-cover of $\mathcal{H}_\infty$.*

Next, we recall a known property of ϵ-covers, which guarantees the success of learning with $\mathcal{G}$.

Lemma 4. *The best-in-class error among $\mathcal{H}_\infty$ and its ϵ-cover $\mathcal{G}$ are at most ϵ apart:*

$$\min_{h\in\mathcal{H}_\infty} R(h) \le \min_{g\in\mathcal{G}} R(g) \le \min_{h\in\mathcal{H}_\infty} R(h) + \epsilon.$$

Proof. The first inequality is trivial since $\mathcal{G}\subseteq\mathcal{H}_\infty$. Let h^* be the best-in-class in $\mathcal{H}_\infty$. By definition, there exists $g^*\in\mathcal{G}$ such that

$$|R(g^*) - R(h^*)| \le \mathcal{L}(g^*,h^*) \le \epsilon \Rightarrow R(g^*) \le R(h^*) + \epsilon.$$

Thus, $\min_{g\in\mathcal{G}} R(g) \le R(g^*) \le R(h^*) + \epsilon$. □

Lemma 4 shows that, when $\mathcal{H}_\infty$ is the family of hypotheses the learner considers, then the approximation error of the ϵ-cover $\mathcal{G}$ is at most ϵ.

Given the minimal ϵ-cover $\mathcal{G}$, the learner can run any active learning algorithm with the finite hypothesis set $\mathcal{H}=\mathcal{G}$ and achieve favorable learning guarantees. For example, we can run IWAL on the minimal ϵ-cover $\mathcal{G}$ and achieve the following results. Let θ be the disagreement coefficient defined with respect to the infinite hypothesis set $\mathcal{H}_\infty$ and its best-in-class predictor h^*.

Theorem 5. *Let $\widehat{g}_T$ be the returned hypothesis after T rounds, For any $\delta>0$, with probability at least $1-\delta$, for all $t>0$,*

$$R(\widehat{g}_T) \le R(h^*) + \epsilon + 2\Delta_T,$$
$$\mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}}[p_t|\mathcal{F}_{t-1}] \le 4\theta\big[R(h^*) + \epsilon/2 + 2\Delta_{t-1}\big],$$

with $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|N(\mathcal{H}_\infty,\epsilon)|^2/\delta)}$.

Proof. Let $g^* = \operatorname{argmin}_{g\in\mathcal{G}} R(g)$. By Theorem 1 and Lemma 4,

$$R(\widehat{g}_T) \le R(g^*) + 2\Delta_T \le R(h^*) + \epsilon + 2\Delta_T.$$

Let $\mathcal{G}_t$ be the version space of IWAL at round t. Then, $\forall g\in\mathcal{G}_t$,

$$R(g) \le R(g^*) + 4\Delta_{t-1} \le R(h^*) + \epsilon + 4\Delta_{t-1}.$$

Thus,

$$\rho(g,h^*) \le R(g) + R(h^*) \le 2R(h^*) + \epsilon + 4\Delta_{t-1}.$$

Let $r_t = 2R(h^*) + \epsilon + 4\Delta_{t-1}$, then $\mathcal{G}_t \subset B(h^*,r_t) = \{h\in\mathcal{H}_\infty : \rho(h,h^*)\le r_t\}$. Thus, by definition of θ,

$$\begin{aligned}&\mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{X}}[p_t|\mathcal{F}_{t-1}]\\ &= \mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}}\Big[\max_{g,g'\in\mathcal{G}_t}\mathcal{L}(g(x_t),g'(x_t))|\mathcal{F}_{t-1}\Big]\\ &\le 2\mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}}\Big[\max_{g\in\mathcal{G}_t}\mathcal{L}(g(x_t),h^*(x_t))|\mathcal{F}_{t-1}\Big]\\ &\le 2\mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}}\Big[\sup_{h\in B(h^*,r_t)}\mathcal{L}(h(x_t),h^*(x_t))|\mathcal{F}_{t-1}\Big]\\ &\le 2\theta r_t = 4\theta\big[R(h^*) + \epsilon/2 + 2\Delta_{t-1}\big].\end{aligned}$$

This completes the proof. □

Theorem 5 shows that by running IWAL on the ϵ-cover $\mathcal{G}$, we can achieve a generalization error that is approaching the best-in-class error among the infinite hypothesis set $\mathcal{H}_\infty$, plus ϵ.

C. More Experimental Results

In Figures 4-6, we show the experimental results for all 8 datasets.

In particular, Figure 6 compares IZOOM with IWAL and PASSIVE learning (which requests all the labels), which are not presented in the main draft due to space limitations.

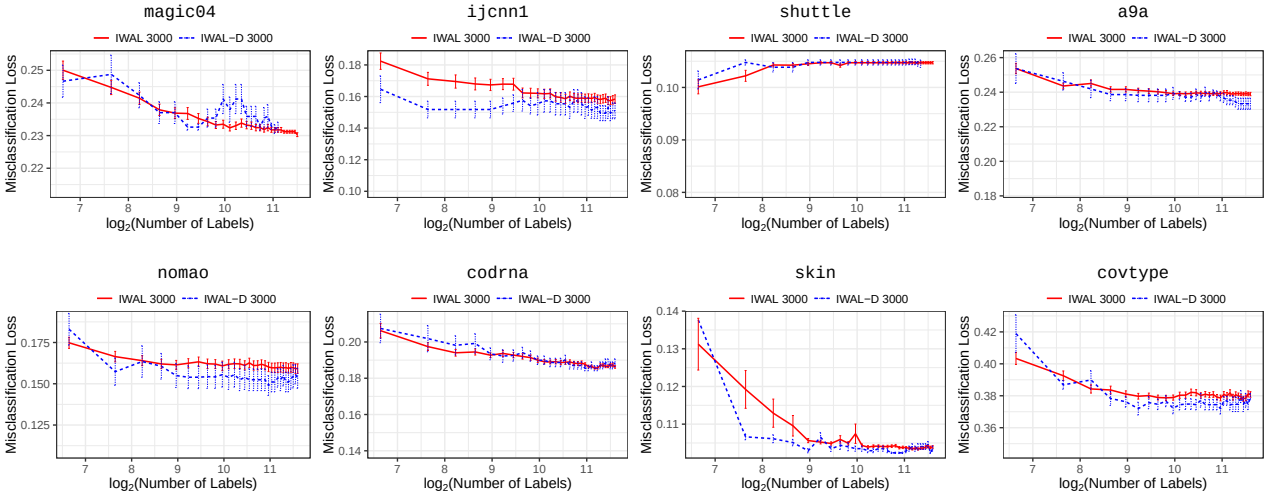

Figure 4. Misclassification loss of IWAL and IWAL-D on hold out test data versus number of labels requested ($\log_2$ scale). In some cases, we observe that IWAL-D achieves a better test performance (`ijcnn1`, `a9a`, `nomao`, `skin`, `covetype`), sometimes mostly at the beginning. In others, the difference is not significant.

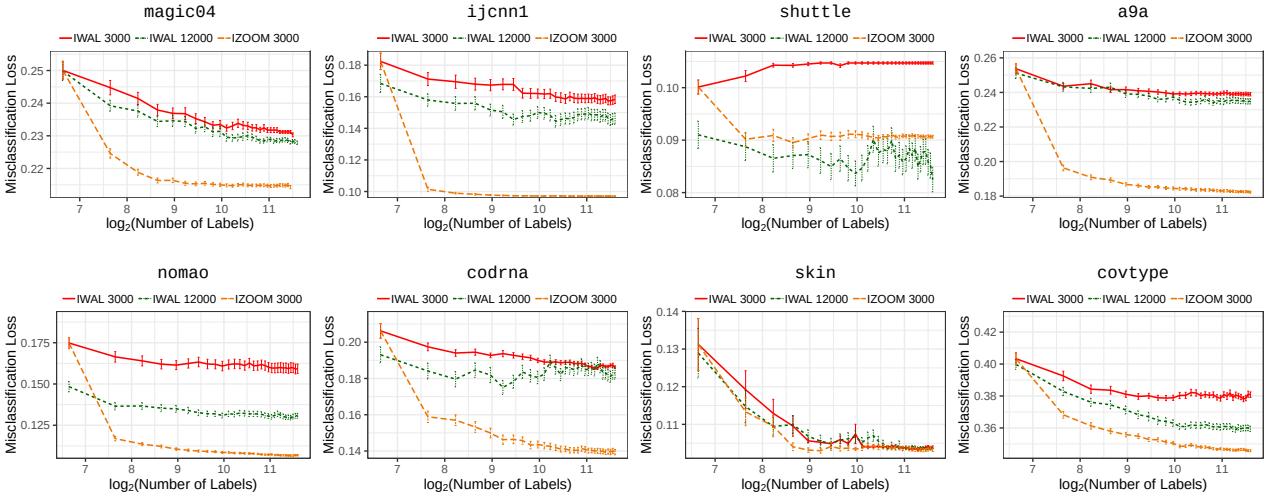

Figure 5. Misclassification loss of IWAL and IZOOM, on hold out test data versus number of labels requested ($\log_2$ scale). The figures show that IZOOM provides significant improvements in performance over IWAL despite it is run with just $|\mathcal{H}| = 3000$. The red curves for IWAL with $|\mathcal{H}| = 3000$ are repetitions from Figure 4.

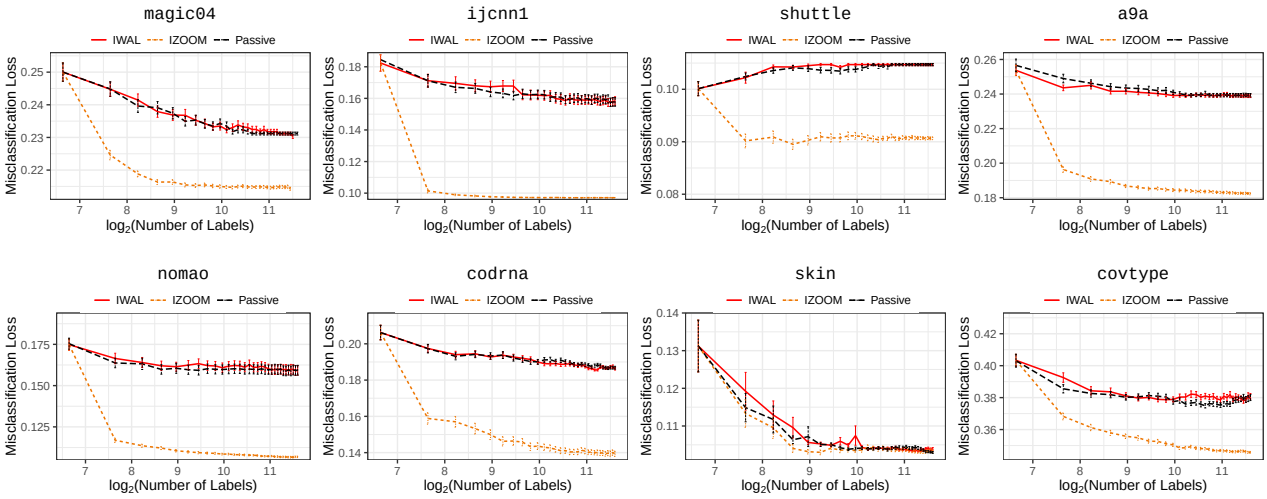

Figure 6. Misclassification loss of IWAL, IZOOM, and PASSIVE learning (request all the labels), all with $|\mathcal{H}| = 3000$, on hold out test data versus number of labels requested ($\log_2$ scale). The figures show that IZOOM provides significant improvements in performance over IWAL and PASSIVE. The red curves for IWAL and orange curves for IZOOM are repetitions from Figure 5.