Region-Based Active Learning

Corinna Cortes
Google Research
New York, NY

Giulia DeSalvo
Google Research
New York, NY

Claudio Gentile
Google Research
New York, NY

Mehryar Mohri
Google Research & Courant
New York, NY

Ningshan Zhang
New York University
New York, NY

Abstract

We study a scenario of active learning where the input space is partitioned into different regions and where a distinct hypothesis is learned for each region. We first introduce a new active learning algorithm (EIWAL), which is an enhanced version of the IWAL algorithm, based on a finer analysis that results in more favorable learning guarantees. Then, we present a new learning algorithm for region-based active learning, ORIWAL, in which either IWAL or EIWAL serve as a subroutine. ORIWAL optimally allocates points to the subroutine algorithm for each region. We give a detailed theoretical analysis of ORIWAL, including generalization error guarantees and bounds on the number of points labeled, in terms of both the hypothesis set used in each region and the probability mass of that region. We also report the results of several experiments for our algorithm which demonstrate substantial benefits over existing non-region-based active learning algorithms, such as IWAL, and over passive learning.

1 Introduction

Standard supervised learning algorithms often rely on large amounts of labeled samples to achieve a high performance. But labeling samples is often very costly since it typically requires human inspection and in some cases high human expertise. Can we learn with a limited labeling budget? This is the challenge of active learning, which remains an active area of research in machine learning, with substantial applications and benefits.

Active learning algorithms seek to request as few labels as possible to learn an accurate predictor. There are two standard settings of active learning: the so-called *pool setting* where the algorithm receives as input an i.i.d. pool of unlabeled points and where it incrementally requests the label of a number of points; and the *on-line setting* where the algorithm receives one i.i.d. point at each round and must decide on whether to request its label. In both cases, after making a number of label requests within a budget, the algorithm returns a predictor chosen out of a hypothesis set, which is hoped to admit a small generalization error. Observe that an active learning algorithm for the on-line setting can also be applied to the pool setting.

In the last few decades, a number of active learning algorithms have been designed, some for specific tasks and requiring strong assumptions. When the problem is separable, Cohn et al. [1994] proposed an algorithm with logarithmic label complexity. A 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 decision boundary. This type of algorithms admits favorable label complexity on the uniform distribution over the unit sphere or on the log-concave distribution. In the pool setting, Dasgupta and Hsu [2008] proposed a hierarchical sampling approach which selectively queries labels from the pool of data and moves down the hierarchies until relatively pure clusters are uncovered. For this type of cluster-based active learning, Urner et al. [2013], Kpotufe et al. [2015] provided a label complexity analysis, but only under various conditions on the data distribution. In the on-line setting, general active learning algorithms [Balcan et al., 2006, Dasgupta et al., 2008, Beygelzimer et al., 2009, 2010, Huang et al., 2015, Zhang and Chaudhuri, 2014] with favorable guarantees both in terms of generalization and label complexity have been devised. These algorithms rely on efficient searching in the concept class, and request labels based on the "disagreement" among hypotheses in the current version space. Their label complexities are bounded in terms of an important quantity known as the disagreement coefficient [Hanneke, 2007]. Among these algorithms, some are computationally inefficient, however, for keeping track of the version space explicitly [Balcan et al., 2006], or for solving expensive optimization problems such as empirical risk minimization with 0-1 loss [Dasgupta et al., 2008, Zhang and Chaudhuri, 2014]. The issue of computational efficiency is one of the key research questions in this area.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

This paper considers the on-line active learning in a novel scenario where the input space is partitioned into a finite number of regions. The problem then consists of requesting labels as in the standard case to learn one predictor for each region. This problem naturally arises in a number of applications, such as speech recognition where the regions are data sources (e.g. broadcast news, conversational speech, email, or dictation), and problems in recommendation systems, where the regions are general categories of an item (e.g., film genres). In all these cases, the regions of the input space are suggested by the application at hand. In other tasks, there may be a natural partitioning into regions based on the features used. Nevertheless, simple partitions of the input space, such as random partitions, are often convenient in the absence of prior knowledge about the nature of the input features, and still provide significant benefit in learning, as empirically shown by our experiments.

In all cases, a different hypothesis set can be used for each region and the hope is that often, but not always, the best-in-class predictor at each region will be very accurate, in fact achieving a loss of almost zero on its region. This is the main motivation for our study of *region-based active learning*. As we shall see, in many applications one can indeed achieve a substantially better performance via this formulation of the problem.

The idea of separating the input space in on-line active learning is novel, as all on-line active learning algorithms available in the literature focus on the standard single region input space. A related area in the pool active learning setting is hierarchical sampling (e.g., [Dasgupta and Hsu, 2008]), where the input space admits a hierarchical clustering structure. This scenario of disjoint input space is partially related to stratified sampling techniques in statistics [Neyman, 1934], where a statistical population is divided into disjoint and homogeneous subgroups. Each subgroup is sampled independently, and different criteria can be used to determine an optimal sample size for each group [Rossi et al., 1983]. One such criterion is the sample variance from an existing sample. While such a strategy will help minimize the overall variance, the technique does not address generalization and comes with no learning guarantees.

In this work, we first introduce a new active learning algorithm (EIWAL), which is an enhanced version of the IWAL algorithm from Beygelzimer et al. [2009], based on a finer analysis that results in more favorable learning guarantees. Then, we present a new learning algorithm for region-based active learning, ORIWAL, in which either IWAL or EIWAL serve as a subroutine. ORIWAL optimally allocates points to the subroutine algorithm for each region. We give a detailed theoretical analysis of ORIWAL, including generalization error guarantees and bounds on the number of points labeled, in terms of both the hypothesis set used in each region and the probability mass of that region. We also report the results of several experiments for our algorithm which demonstrate substantial benefits over existing non-region-based active learning algorithms, such as IWAL, and over passive learning.

The rest of this paper is organized as follows. Section 2 introduces the definitions and notation needed for our analysis and specifies the learning scenario we consider. In Section 3, we introduce the EIWAL algorithm, and prove the associated theoretical guarantees. Section 4 presents our novel region-based active learning algorithm ORIWAL and its learning guarantees. In Section 5, we report the results of our experiments in several datasets. Section 6 concludes the paper with a discussion of future work.

2 Preliminaries

In this section, we first introduce the definitions and notation relevant to our analysis and next describe the active learning scenario we consider.

Definitions. We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ the input space and by $\mathcal{Y} = \{-1, +1\}$ the binary output space. We assume given a partitioning of $\mathcal{X}$ into n disjoint regions: $\mathcal{X} = \bigcup_{k=1}^n \mathcal{X}_k$, with $\mathcal{X}_k \cap \mathcal{X}_{k'} = \emptyset$ for $k \neq k'$. This partitioning may have been generated at random or selected in some other way based on some prior knowledge about the task. In all cases, it is assumed to be fixed before receiving sample points.

As in standard supervised learning, we assume that training and test points are drawn i.i.d. according to some unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We will denote by $\mathsf{p}_k = \mathbb{P}(\mathcal{X}_k)$ the probability mass of region $\mathcal{X}_k$ with respect to the marginal distribution induced by $\mathcal{D}$ over $\mathcal{X}$. For each $k \in [n]$, we denote by $\mathcal{H}_k$ the hypothesis set used for region $\mathcal{X}_k$, which consists of functions mapping from $\mathcal{X}$ to some prediction space $\mathcal{Z} \subseteq \mathbb{R}$. In the simplest case, the same hypothesis set is chosen for all regions: $\mathcal{H}_1 = \cdots = \mathcal{H}_n$.

We denote by $\ell \colon \mathcal{Z} \times \mathcal{Y} \to [0, 1]$ the loss function. The loss function we adopt in the implementations run in our experiments is the standard logistic loss, defined for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and hypothesis h by $\log(1 + e^{-yh(x)})$, which we then normalize to be in $[0, 1]$. We will denote by $R(h)$ the generalization error or expected loss of a hypothesis h: $R(h) = \mathbb{E}[\ell(h(x), y)]$. Similarly, for any $k \in [n]$, we denote by $R_k(h)$ the expected loss of h on region $\mathcal{X}_k$: $R_k(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y) \mid x \in \mathcal{X}_k]$. Thus, for any hypothesis h, we have $R(h) = \sum_{k=1}^n \mathsf{p}_k R_k(h)$.

We will denote by $\mathcal{H}_{[n]}$ be the set of aggregate region-based hypotheses:

$$\mathcal{H}_{[n]} = \left\{ \sum_{k=1}^n 1_{x \in \mathcal{X}_k} h_k(x) \colon h_k \in \mathcal{H}_k \right\},$$

whose size $|\mathcal{H}_{[n]}|$ equals $\prod_{k=1}^n |\mathcal{H}_k|$. We denote by h^* the best-in-class hypothesis in $\mathcal{H}_{[n]}$, that is, $h^* =$

$\operatorname{argmin}_{h \in \mathcal{H}_{[n]}} R(h)$, and similarly denote by h_k^* the best-in-class hypothesis in region $\mathcal{X}_k$: $h_k^* = \operatorname{argmin}_{h \in \mathcal{H}_k} R_k(h)$. For simplicity, we denote by $R^* = R(h^*)$ and $R_k^* = R_k(h_k^*)$ the error of overall and region-specific best-in-class, respectively. The best-in-class hypothesis $h^* \in \mathcal{H}_{[n]}$ can be expressed as follows in terms of the h_k^*s:

$$\begin{aligned} h^*(x) &= \operatorname*{argmin}_{h \in \mathcal{H}_{[n]}} \sum_{k=1}^{n} \mathrm{p}_k R_k(h) \qquad (1) \\ &= \sum_{k=1}^{n} 1_{x \in \mathcal{X}_k} \left[\operatorname*{argmin}_{h \in \mathcal{H}_k} R_k(h) \right] = \sum_{k=1}^{n} 1_{x \in \mathcal{X}_k} h_k^*(x) \, . \end{aligned}$$

Observe, however, that the risk minimization over each region individually is always more advantageous than the risk minimization over the entire space, for the minimal error within the aggregate region-based hypothesis set $\mathcal{H}_{[n]}$ is always less than or equal to the minimal error achieved by selecting each single hypothesis for all regions. Too see this, consider the simplest case where $\mathcal{H}_1 = \cdots = \mathcal{H}_n = \mathcal{H}$. Then, by the super-additivity of the min operator, the following holds:

$$\begin{aligned} R(h^*) &= \sum_{k=1}^{n} \mathrm{p}_k \left[\min_{h \in \mathcal{H}} R_k(h) \right] \\ &\leq \min_{h \in \mathcal{H}} \left[\sum_{k=1}^{n} \mathrm{p}_k R_k(h) \right] = \min_{h \in \mathcal{H}} R(h) \, . \end{aligned}$$

In other words, the approximation error of $\mathcal{H}_{[n]}$ is always less than or equal to that of $\mathcal{H}$, implying that $\mathcal{H}_{[n]}$ is always significantly richer than any individual hypothesis set $\mathcal{H}$.

Learning scenario. We consider active learning in the *on-line setting*. Unlike the *pool-based setting* where the learner receives the full set of unlabeled points beforehand, in the on-line setting, at each round $t \in [T] = \{1, \ldots, T\}$, the learner receives a point x_t drawn i.i.d. according to the marginal distribution induced by $\mathcal{D}$ on $\mathcal{X}$. She then either selects to request the label of x_t, in which case she receives its label y_t, or chooses not to solicit x_t's label.

The quality of an active learning algorithm is measured by two quantities in this setting: the generalization error of the hypothesis $h \in \mathcal{H}_{[n]}$ it returns after the T rounds, and the number of labels it requests after T rounds.

3 Enhanced-IWAL Algorithm

In this section, we present an enhanced version of the IWAL (Importance Weighted Active Learning) algorithm of Beygelzimer et al. [2009], called EIWAL.

Algorithms such as IWAL use importance weights to address key the issue of sampling bias in active learning. Beygelzimer et al. [2009] gave theoretical guarantees both for the generalization error and the label complexity of IWAL. Our enhanced version of IWAL admits improved confidence intervals, and thus sharper performance guarantees than the original IWAL, especially in the case where the best-in-class error $R(h^*)$ is small. In that small error regime, EIWAL also improves upon a more recent and more refined importance-weighted active learning algorithm discussed in Beygelzimer et al. [2010] (Theorem 3 therein). This advantage is particularly significant in the scenario of region-based active learning that we are interested in where, often with a large number of regions, the region-based best-in-class errors $R_k(h_k^*)$ are small.

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

$$p_t = \max_{f, g \in \mathtt{H}_t \; y \in \mathcal{Y}} \ell(f(x_t), y) - \ell(g(x_t), y) \, .$$

If $Q_t = 1$, then the label y_t is requested and $\mathtt{H}_t$ is trimmed to $\mathtt{H}_{t+1}$ via an importance-weighted empirical risk minimization:

$$\mathtt{H}_{t+1} = \left\{ h \in \mathtt{H}_t \colon \frac{1}{t} \sum_{s=1}^{t} \frac{Q_s}{p_s} \ell(h(x_s), y_s) \leq L_t^* + \Delta_t \right\},$$

where L_t^* is given by

$$L_t^* = \min_{h \in \mathtt{H}_t} \frac{1}{t} \sum_{s=1}^{t} \frac{Q_s}{p_s} \, \ell(h(x_s), y_s) \, ,$$

and where the slack term Δ_t is of the form[1]

$$\frac{1}{t} \left[\sqrt{ \left[\sum_{s=1}^{t} p_s \right] \log \left[\frac{t|\mathcal{H}|}{\delta} \right] } + \log \left[\frac{t|\mathcal{H}|}{\delta} \right] \right] .$$

The definition of the slack term Δ_t is the main significant difference between EIWAL and the original IWAL. In the latter, $(\sum_{s=1}^{t} p_s)$ is replaced by the crude upper bound t: $\Delta_t = \frac{1}{t} \sqrt{t \, \log(t|\mathcal{H}|/\delta)}$. The final hypothesis h_T returned by EIWAL is defined as in IWAL:

$$h_T = \operatorname*{argmin}_{h \in \mathtt{H}_T} \frac{1}{T} \sum_{t=1}^{T} \frac{Q_t}{p_t} \, \ell(h(x_t), y_t) \, .$$

For our theoretical analysis of EIWAL, we will adopt the definitions and concepts in Beygelzimer et al. [2009]. Define the distance between two hypotheses $f, g \in \mathcal{H}$ as

$$\rho(f, g) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \max_{y} \left| \ell(f(x), y) - \ell(g(x), y) \right| \, .$$

[1] See the exact expression in the proof of Theorem 1 in Appendix A.

Given $r > 0$, let $B(f, r)$ denote the ball of radius r centered in $f \in \mathcal{H}$: $B(f, r) = \{g \in \mathcal{H} \colon \rho(f, g) \leq r\}$. The generalized disagreement coefficient $\theta(\mathcal{D}, \mathcal{H})$ can then be defined as follows:

$$\theta(\mathcal{D}, \mathcal{H}) = \inf_{\theta} \Big\{ \forall r \geq 0, \\ \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \sup_{h \in B(h^*, r)} \sup_{y} |\ell(h(x), y) - \ell(h^*(x), y)| \leq \theta r \Big\},$$

where $h^* = \mathrm{argmin}_{h \in \mathcal{H}} R(h)$. The disagreement coefficient θ is a complexity measure widely used in disagreement-based active learning problems. In particular, Hanneke [2007] proved upper and lower bounds for the label complexity for the A^2 algorithm in terms of the disagreement coefficient θ. Dasgupta et al. [2008] also gave an upper bound for the DHM algorithm using θ. See [Hanneke, 2014] for a more extensive analysis of the disagreement coefficient and active learning.

Using the definitions and concepts just introduced, the following theoretical guarantees can be proven for EIWAL.[2]

Theorem 1 (EIWAL)**.** *Let h_T denote the hypothesis returned by* EIWAL *after T rounds and τ_T the total number of requested labels. Then, for all $\delta > 0$, with probability at least $1 - \delta$, for any $T > 0$ the following inequality holds:*

$$R(h_T) \leq R(h^*) + \frac{2}{T} \left[\sqrt{\textstyle\sum_{t=1}^T p_t} + 6 \sqrt{\log \left[\frac{2(3+T)T^2}{\delta} \right]} \right] \\ \times \sqrt{\log \left[\frac{16 T^2 |\mathcal{H}|^2 \log(T)}{\delta} \right]}.$$

Moreover, with probability at least $1 - \delta$, for any $T > 0$, the following inequality holds:

$$\tau_T \leq 8\theta K_l \left(R(h^*)T + O(\sqrt{R(h^*)T \log(T|\mathcal{H}|/\delta)}) \right) \\ + O(\log^3(T|\mathcal{H}|/\delta)),$$

where K_ℓ is a constant that depends on the loss function ℓ.

For reference, the generalization bound given in [Beygelzimer et al., 2009] for IWAL admits the following form:

$$R(h_T) \leq R(h^*) + \sqrt{\frac{1}{T} \log \left[\frac{T^2 |\mathcal{H}|^2}{\delta} \right]}, \qquad (2)$$

and the bound on the number of labels is given by

$$\tau_T = O\left(\theta K_l \left(R(h^*)T + \sqrt{T \log(T|\mathcal{H}|/\delta)} \right) \right). \qquad (3)$$

The comparison of Theorem 1 with (2) and (3), as well as with Beygelzimer et al. [2010] (Theorem 3 therein) shows the following: the bound on the generalization error $R(h_T)$ in Theorem 1 is at least as favorable as the $1/\sqrt{T}$ rate of these previous results, since $\sum_{t=1}^T p_t \leq T$. Furthermore, the bound on the number of labels τ_T is better than both (3) and the results in Beygelzimer et al. [2010] when $R(h^*)$ is small, since we have an extra $R(h^*)$ inside the square root. In fact, in the separable case where $R(h^*) = 0$, our label complexity bound is $\log^3(T)$, which is only polylogarithmic in T, as opposed to the $\sqrt{T}$ guarantee of both (3) and Beygelzimer et al. [2010]. Similarly, when $R(h^*) = 0$, one can see that the generalization error bound of EIWAL has the form $\log^2(T)/T$, rather than $1/\sqrt{T}$ of (2) and Beygelzimer et al. [2010]. This is because $\sum_{t=1}^T p_t$ concentrates fast around τ_T which, as we just said, is only $O(\log^3(T))$ when $R(h^*) = 0$.

Algorithm 1 ORIWAL$((\mathcal{H}_k)_{k \in [n]}, (\mathrm{p}_k)_{k \in [n]}, \delta, T)$

```
for $k \in [n]$ do
    $c_k \leftarrow \log \left[ \frac{16T^2|\mathcal{H}_k|^2 \log(T) n}{\delta} \right]$
    $\alpha_k \leftarrow \frac{(c_k/\mathrm{p}_k)^{\frac{1}{3}}}{\max_{k \in [n]} (c_k/\mathrm{p}_k)^{\frac{1}{3}}}$
end for
for $t \in [T]$ do
    RECEIVE$(x_t)$
    $k_t \leftarrow k$ such that $x_t \in \mathcal{X}_k$
    $B \sim$ Bernoulli$(\alpha_{k_t})$
    if $B = 1$ then
        $h_{k,t} \leftarrow$ EIWAL$_{k_t}(x_t)$
        Request $y_t$ according to EIWAL$_{k_t}$ on input $x_t$
        Update (if any) internal state of EIWAL$_{k_t}$
    end if
end for
Return $h_T \leftarrow \left[ x \mapsto \sum_{k=1}^n 1_{x \in \mathcal{X}_k} h_{k,T}(x) \right]$
```

4 Region-Based Active Learning

In this section, we describe an active learning algorithm, ORIWAL (Optimal Region-based IWAL), under the region-based setting. The algorithm works by running a separate subroutine EIWAL on each of the n regions, while carefully allocating labeling resources across the regions.

4.1 The ORIWAL Algorithm

At each time t, ORIWAL receives an unlabeled point x_t that belongs to region $\mathcal{X}_{k_t}$, for some $k_t \in [n]$. Then, with some probability α_k, ORIWAL decides whether to send x_t to subroutine EIWAL$_{k_t}$, the EIWAL algorithm running on region $\mathcal{X}_{k_t}$. If x_t is sent to EIWAL$_{k_t}$, then it is EIWAL$_{k_t}$ that determines whether to request the associated label y_t. Thus, y_t is requested only if x_t is passed to EIWAL$_{k_t}$ (probability α_k) *and* EIWAL$_{k_t}$ happens to ask for this label (probability p_t depending on the current state of EIWAL$_{k_t}$). The pseudocode of ORIWAL is given in Algorithm 1.

[2] Due to space limitations, the proofs of all our main results are given in the appendix.

In what follows, when ORIWAL passes x_t to EIWAL_{k_t}, we say that ORIWAL *queries* EIWAL_{k_t}. Notice that querying EIWAL_k to determine whether to ask for a label is computationally much more expensive than determining whether or not to pass a point to EIWAL_k. Thus, we will discuss the learning guarantees and label complexity bounds of ORIWAL in terms of the number of queries to the EIWAL subroutines.

The crux of the ORIWAL algorithm rests on finding the probabilities α_k, which determine how many points in expectation are passed to the k-th region, so as to optimize learning guarantees. Ideally, the algorithm should not pass points to a region where the subroutine has already found a good hypothesis. Regions in need for labels are those where the corresponding subroutines have received few points or where a larger number of points is needed to identify an accurate hypothesis.

To determine the probabilities α_k, we first use the theoretical guarantees derived for EIWAL to determine T_k, the number of queries made to EIWAL_k operating in region $\mathcal{X}_k$. At a high level, the optimal setting of T_ks, which translates into an optimal setting of α_ks, is one that admits the best trade-off between generalization guarantee and label complexity bound. By Theorem 1, the generalization bound of EIWAL_k is proportional to a complexity term c_k of the form [3] $c_k = \log\left[\frac{16T^2|\mathcal{H}_k|^2\log(T)n}{\delta}\right]$, where we upper bound $\log T_k$ by $\log T$, and further upper bound all label requesting probabilities p_t by 1. [4] Hence, to determine the optimal setting of T_ks, we need to find $T_1, T_2, \ldots, T_n$ satisfying:

$$\min_{T_1,\cdots,T_n} \sum_{k=1}^{n} \mathrm{p}_k \sqrt{\frac{c_k}{T_k}}, \text{ s.t. } \sum_{k=1}^{n} T_k \leq T\,,$$

where $\mathrm{p}_k = \mathbb{P}(\mathcal{X}_k)$. It is straightforward to show that the optimal solution T_k^* admits the following form:

$$T_k^* = \left[\frac{\mathrm{p}_k^{\frac{2}{3}} c_k^{\frac{1}{3}}}{\sum_{k'=1}^{n} \mathrm{p}_{k'}^{\frac{2}{3}} c_{k'}^{\frac{1}{3}}}\right] T.$$

We then choose the probabilities α_ks such that, given the total number T of possible queries, the expected number of queries to EIWAL_k matches T_k^*. That is, α_k should satisfy

$$\frac{\mathrm{p}_k\alpha_k}{\sum_{k'=1}^{n} \mathrm{p}_{k'}\alpha_{k'}} = \frac{T_k^*}{T} = \frac{\mathrm{p}_k^{\frac{2}{3}} c_k^{\frac{1}{3}}}{\sum_{k'=1}^{n} \mathrm{p}_{k'}^{\frac{2}{3}} c_{k'}^{\frac{1}{3}}}\,, \tag{4}$$

where the left-most side is the conditional probability of querying EIWAL_k, conditioning on a total number T of queries, and the right-most side is the optimal allocation proportion determined by T_k^*. It is straightforward to show that for any $\lambda > 0$, $\alpha_k = \lambda(c_k/\mathrm{p}_k)^{\frac{1}{3}}$ would satisfy (4).

Finally, to determine the optimal setting of α_ks, we need to determine the last parameter λ. Observe that, for a given λ and its corresponding α_ks, a total of $\sum_{k=1}^{n} \mathrm{p}_k\big(1-\alpha_k\big)$ unlabeled points will be discarded due to the "**if** $B = 1$ **then** ..." step of ORIWAL (Algorithm 1). Thus, we choose λ that minimizes the number of discarded unlabeled points:

$$\min_{\lambda\geq 0} \sum_{k=1}^{n} \mathrm{p}_k\big(1-\lambda(c_k/\mathrm{p}_k)^{\frac{1}{3}}\big), \text{ s.t. } \lambda\big(c_k/\mathrm{p}_k\big)^{\frac{1}{3}} \leq 1, \forall k \in [n].$$

The constraint on λ ensures that α_ks are valid probabilities: $\alpha_k \leq 1$, $\forall k \in [n]$. Solving the above problem yields the optimal setting of α_ks:

$$\lambda = \frac{1}{\max_{k\in[n]}(c_k/\mathrm{p}_k)^{\frac{1}{3}}}, \quad \alpha_k = \frac{(c_k/\mathrm{p}_k)^{\frac{1}{3}}}{\max_{k\in[n]}(c_k/\mathrm{p}_k)^{\frac{1}{3}}}. \tag{5}$$

Observe that in the expression of α_ks, we assumed access to the probability mass p_k of each region. This is a reasonable assumption in many applications of active learning, since accurately estimating p_k only requires unlabeled data. Hence, we can conceive a preprocessing stage where the probabilities p_k are accurately estimated from large amounts of unlabeled data. Alternatively, these probabilities can be estimated incrementally, and our analysis can be extended to cover that way of proceeding as well.

4.2 Theoretical Analysis

For α_ks defined as in (5), the following theoretical guarantees hold for the returned hypothesis and label complexity. The guarantees of ORIWAL depend on region-based disagreement coefficient $\theta_k = \theta(\mathcal{D}_k, \mathcal{H}_k)$, where $\mathcal{D}_k = \mathcal{D}|\mathcal{X}_k$ is defined as the conditional distribution of x on region k.

Theorem 2. *For any $\delta > 0$, with probability at least $1 - \delta$, for any $T > 0$, the following inequality holds for the hypothesis returned by* ORIWAL *(Algorithm 1) at time T:*

$$\begin{aligned} R(h_T) \leq R(h^*) &+ \sum_{k=1}^{n} 2\mathrm{p}_k \sqrt{\frac{4\theta_k K_\ell R_k^*}{T_k} \log\left[\frac{16T_k^2|\mathcal{H}_k|^2\log(T_k)n}{\delta}\right]} \\ &+ \Big(\sum_{k=1}^{n} \frac{\mathrm{p}_k}{T_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T_k|\mathcal{H}_k|n/\delta\big)\Big), \end{aligned}$$

where T_k is number of queries made to IWAL_k*. Moreover, with probability at least $1-\delta$, for any $T > 0$, the following inequality holds for the number of requested labels τ_T:*

$$\begin{aligned} \tau_T \leq \sum_{k=1}^{n} \Big(& 8\theta_k K_l \Big[R_k^* T_k + O\big(\sqrt{R_k^* T_k \log(T_k|\mathcal{H}_k|n/\delta)}\big)\Big] \\ &+ O(\log^3(T_k|\mathcal{H}_k|n/\delta))\Big). \end{aligned}$$

The generalization bound is the sum of the generalization error of the best in class $h^* \in \mathcal{H}_{[n]}$ and the sum of the

[3] The extra factor n is due to a union bound over the n regions, so as to make Theorem 1 hold for all regions simultaneously.

[4] In Section 4.4, we will present the version of ORIWAL derived from using the original requesting probabilities p_t.

complexity terms of the hypothesis sets $\mathcal{H}_k$. In particular, if the probability mass p_k of region $\mathcal{X}_k$ is small, then the corresponding complexity term of set $\mathcal{H}_k$ is given less weight. Moreover, as one could expect, the overall bound becomes tighter as the number of queries T_k made to EIWAL$_k$ increases. For the label complexity bound of τ_T, the term inside the bracket is of the same form as the term in the label complexity bound of EIWAL for a single region. In this case, the region-specific disagreement coefficients θ_k, best-in-class error R_k^*, and complexity terms $\log|\mathcal{H}_k|$, scale the contribution of each region accordingly.

We can also derive guarantees that do not depend on the empirical quantities T_k, but only on T. When the expected number of passed samples per region is at least $O(\log n)$, we have the following result. For sake of brevity, we denote by q_k the optimal allocation proportion in Equation (4):

$$\mathsf{q}_k = \frac{\mathsf{p}_k^{2/3} c_k^{1/3}}{\sum_{k'=1}^n \mathsf{p}_{k'}^{2/3} c_{k'}^{1/3}}, \qquad k \in [n] \,.$$

Corollary 3. *For all $\delta > 0$, with probability at least $1-\delta$, for any $T \geq \frac{4\log(2n/\delta)}{\min_{k\in[n]}\mathsf{q}_k}$ the following inequality holds:*

$$\begin{aligned} & R(h_T) \\ & \leq R(h^*) + 2\sum_{k=1}^n \mathsf{p}_k \sqrt{\frac{4\theta_k K_l R_k^*}{T\mathsf{q}_k}\log\left[\frac{32T^2|\mathcal{H}_k|^2\log(T)n}{\delta}\right]} \\ & + \Big(\sum_{k=1}^n \frac{\mathsf{p}_k}{T\mathsf{q}_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T|\mathcal{H}_k|n/\delta\big)\Big) \,. \end{aligned}$$

Moreover, with probability at least $1-2\delta$, for all $T > 0$, the following inequality holds:

$$\begin{aligned} \tau_T \leq 8K_\ell & \Big[\sum_{k=1}^n \theta_k R_k^* T\mathsf{q}_k\Big] \\ & + \sum_{k=1}^n O\Big(\sqrt{R_k^* T\mathsf{q}_k \log\Big[\frac{T|\mathcal{H}_k|n}{\delta}\Big]}\Big) \\ & + O\Big(n\log^3\big(T\max_{k\in[n]}|\mathcal{H}_k|n/\delta\big)\Big) \,. \end{aligned}$$

We have been discussing the learning guarantees in terms of the number of queries to the EIWAL subroutines, and we do not take into account the number of rounds in which the ORIWAL decides not to query EIWAL. This is because, as we have mentioned earlier, querying the EIWAL subroutine consumes a significantly larger amount of computational resources than determining whether to make a query. This view of the learning problem naturally arises in applications where the unlabeled samples are inexpensive and are processed beforehand, so it takes no time to determine their regions and to sample a Bernoulli random variable to decide whether to query. In other words, given a limited amount of resources, we are more interested in the performance of the algorithm in terms of the number of expensive operations, i.e., queries to the subroutines, than in terms of the number of rounds where no expensive operations are made.

4.3 Discussion

The advantage of ORIWAL over non-region-based algorithms is twofold: it seeks region-specific best-in-class hypotheses, and it controls the number of queries on each region in an optimal way. If ORIWAL does not optimize for query allocations but instead sets $\alpha_k = 1$ for all $k \in [n]$, ORIWAL reduces to a special region-based algorithm we call RIWAL (Region-based IWAL). RIWAL still enjoys the advantage of region-based hypotheses, but it simply passes on all the points to the subroutines. The only algorithmic difference between ORIWAL and RIWAL is that the former generates a Bernoulli random variable for each incoming sample point, which only consumes a negligible amount of time compared to querying subroutine EIWAL. Given the same number of queries to EIWAL, the two algorithms therefore have comparable computational cost.

Yet, the learning guarantee of ORIWAL is potentially more favorable than that of RIWAL, since ORIWAL explicitly optimizes for the allocations T_k among a fixed budget of T queries to EIWAL. Given a total of T queries, Corollary 3 provides the generalization error of the hypothesis returned after T rounds, in terms of $\mathsf{q}_k = \mathsf{p}_k\alpha_k/(\sum_{k'}^n \mathsf{p}_{k'}\alpha_{k'})$, the probability of querying EIWAL$_k$, conditioned on a query being made. Upper bounding the constants $4\theta_k K_\ell R_k^*$ by 1 gives the following:

$$R(h_T^{\text{RIWAL}}) \leq R(h^*) + 2\sum_{k=1}^n \mathsf{p}_k\sqrt{\frac{c_k}{T\mathsf{q}_k^{\text{RIWAL}}}} + O\Big(\frac{\mathsf{p}_k}{T\mathsf{q}_k^{\text{RIWAL}}}\Big),$$

$$R(h_T^{\text{ORIWAL}}) \leq R(h^*) + 2\sum_{k=1}^n \mathsf{p}_k\sqrt{\frac{c_k}{T\mathsf{q}_k^{\text{ORIWAL}}}} + O\Big(\frac{\mathsf{p}_k}{T\mathsf{q}_k^{\text{ORIWAL}}}\Big).$$

RIWAL sets $\alpha_k = 1$ and thus $\mathsf{q}_k^{\text{RIWAL}} = \mathsf{p}_k$. Meanwhile, by definition, $\mathsf{q}_k^{\text{ORIWAL}} = \mathsf{p}_k^{2/3}c_k^{1/3}/(\sum_{k'=1}^n \mathsf{p}_{k'}^{2/3}c_{k'}^{1/3})$. Disregarding lower order terms, i.e., the third term in the two upper bounds above, the application of Jensen's inequality to the convex function $x \mapsto x^{\frac{3}{2}}$ yields

$$\sum_{k=1}^n \mathsf{p}_k\sqrt{\frac{c_k}{T\mathsf{q}_k^{\text{ORIWAL}}}} = \sqrt{\frac{1}{T}}\Bigg[\sum_{k=1}^n \mathsf{p}_k\Big[\frac{c_k}{\mathsf{p}_k}\Big]^{\frac{1}{3}}\Bigg]^{\frac{3}{2}} \tag{6}$$

$$\leq \sqrt{\frac{1}{T}}\Bigg[\sum_{k=1}^n \mathsf{p}_k\Big[\frac{c_k}{\mathsf{p}_k}\Big]^{\frac{1}{2}}\Bigg] = \sum_{k=1}^n \mathsf{p}_k\sqrt{\frac{c_k}{T\mathsf{q}_k^{\text{RIWAL}}}}. \tag{7}$$

Thus ORIWAL yields a potentially more favorable learning guarantee than RIWAL given the same number of T queries to subroutines. Note that the potential improvement of ORIWAL over RIWAL, that is the difference between (6) and (7), depends on how the ratios c_k/p_k vary across regions. Unbalanced ratio values across regions make (6) significantly smaller than (7), while in the case where c_k/p_k coincide for all $k \in [n]$, there is no improvement.

4.4 ORIWALwith time-varying α_k

When deriving the optimal value of α_ks, we upper bounded $\sum_{t=1}^{T} p_t$ by T in order to simplify the discussion, but there is a finer analysis based on a tighter bound on the complexity term, which results in finding an optimal time-varying $\alpha_k(t)$. Without upper bounding this sum of probabilities, the complexity term of Theorem 1 is $C_k(T_k) = c_k\beta_k(T_k)$, where $\beta_k(T_k) = \left(\sum_t p_t 1_{x_t \in \mathcal{X}_k}\right)/T_k$ is the label requesting probability on region k, averaged over the T_k queries. Note that $C_k(T_k) \leq c_k$. Now, since $C_k(T_k)$ depends on T_k which is an unknown quantity at any given round $t \in [T]$, we cannot directly use it to solve the optimization problem:

$$\min_{T_1,\cdots,T_n} \sum_{k=1}^{n} \mathrm{p}_k \sqrt{\frac{C_k(T_k)}{T_k}}, \text{ s.t. } \sum_{k=1}^{n} T_k \leq T\,.$$

However, by definition, the label requesting probabilities of EIWAL are non-increasing, which implies that $\beta_k(T_k)$ as well as $C_k(T_k)$ are also non-increasing. Thus, at a given current round $t \in [T]$, we can upper bound the above optimization problem by

$$\min_{T_k \geq t_k, k\in[n]} \sum_{k=1}^{n} \mathrm{p}_k \sqrt{\frac{C_k(t_k)}{T_k}}\ \text{, s.t. } \sum_{k=1}^{n} T_k \leq T\,,$$

where t_k denotes the number of queries made for region k at a time t. Via a similar reasoning as before, the solution of this optimization problem leads to setting $\alpha_k(t_k) = \frac{(C_k(t_k)/\mathrm{p}_k)^{\frac{1}{3}}}{\max_{k\in[n]}(C_k(t_k)/\mathrm{p}_k)^{\frac{1}{3}}}$. ORIWAL therefore uses these time-varying quantities $\alpha_k(t_k)$ at each time t instead of α_k in Algorithm 1 to determine whether to query EIWAL$_k$.

By using the time-varying and algorithm-dependent quantities $C_k(t_k)$, ORIWAL gains more information about the current state of each region, and uses it to more efficiently allocate labeling resources. More concretely, according to Lemma 6 in Appendix A, $\beta_k(t_k) = 4\theta K_l R_k^* + O(\sqrt{R_k^*/t_k})$. Thus, when $C_k(t_k)/\mathrm{p}_k$ is relatively large for region k (which implies that $\alpha_k(t_k)$ is relatively large), then either t_k is small and $O(\sqrt{R_k^*/t_k})$ is large, so that EIWAL$_k$ is still learning, or t_k is large but the best-in-class error scaled by the probability of that region, R_k^*/p_k, is large. In both cases, ORIWAL allocates more weight to this region, which needs more labeling resources to learn. In the experiments, we ran ORIWAL with the time-varying $\alpha_k(t_k)$s.

Finally, in Appendix C, we present another extension of IWAL to the region-based setting, called NAIVE-IWAL, which simply runs IWAL with the composite hypothesis set $\mathcal{H}_{[n]}$. We show that NAIVE-IWAL is less favorable in terms of theoretical guarantees than RIWAL, thus is less favorable than ORIWAL as well.

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 observations.

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

5 Experiments

In this section, we report the results of experiments comparing the ORIWAL, RIWAL, and IWAL algorithms. We also compared these active learning algorithms with two passive learning algorithms: PASSIVE, which simply requests the label for all points and finds the hypothesis with the smallest empirical logistic loss, and RPASSIVE, which runs PASSIVE on each region separately.

We experimented with the algorithms just mentioned in 8 binary classification datasets from the UCI repository: magic04, nomao, shuttle, a9a, ijcnn1, codrna, skin, covtype. Table 1 gives summary statistics for these 8 datasets. Note that, for each dataset, we kept the first 10 principal components of the original features. For each dataset, we randomly shuffled the data and ran the algorithms on the first 50% of the data, and tested the learned classifier on the remaining 50%. This was repeated 50 times on each dataset, and the results were averaged.

We randomly drew 3,000 hyperplanes with bounded norms as our base hypothesis set, which we call $\mathcal{H}$, and used these 3,000 hyperplanes as $\mathcal{H}_k$ for all regions $\mathcal{X}_k$, thus, we chose $\mathcal{H}_k = \mathcal{H}$ for all $k \in [n]$. To generate disjoint regions, for each dataset we constructed random binary trees, i.e., binary trees with random splitting criteria, and used the resulting terminal nodes as the disjoint regions. Note that these regions are generated without using any labels.

Below, we present these results for four datasets with 10 disjoint regions. The results for the remaining datasets, as well as for the case where we instead have 20 disjoint regions are provided in Appendix D. In Appendix D, we also contrast the performance of ORIWAL with 10 regions vs. ORIWAL with 20 regions.

We first compared the two region-based active learning algorithms, RIWAL and ORIWAL, and the region-based passive learning algorithm RPASSIVE. Both RIWAL and RPASSIVE were run with the same regions and hypothesis sets as ORIWAL, thus all three algorithms have the same model complexity. Figure 1 plots the misclassification loss on held-out

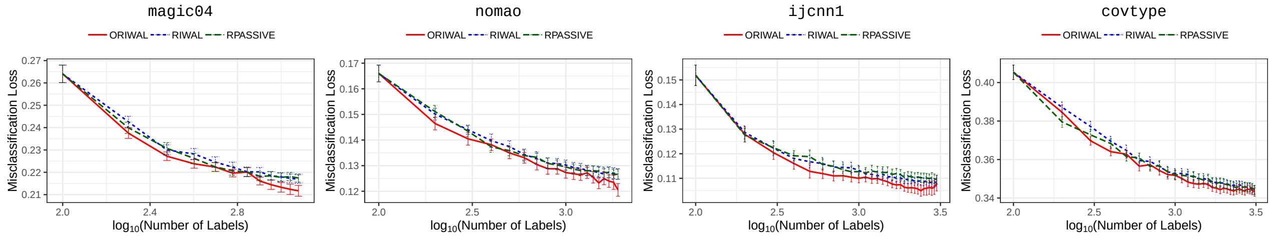

Figure 1: Misclassification loss of ORIWAL, RIWAL, and RPASSIVE on hold out test data versus number of labels requested ($\log_{10}$ scale). The input space was divided into 10 regions. The figures show that ORIWAL typically has a lower misclassification loss than RIWAL and RPASSIVE.

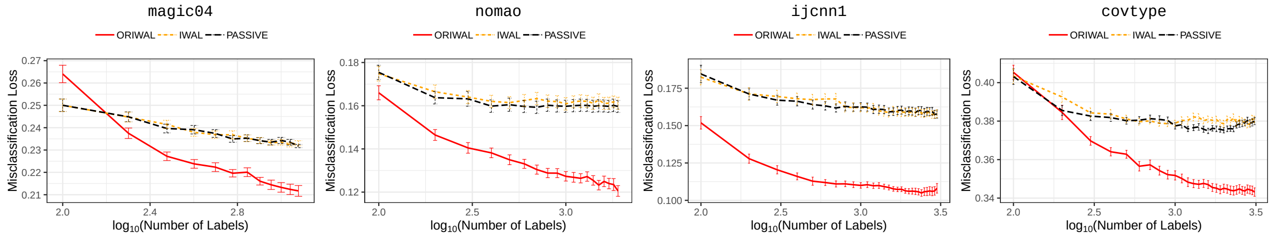

Figure 2: Misclassification loss of ORIWAL (our algorithm), non-region-based IWAL, and non-region-based passive learning PASSIVE on held-out test data, plotted as a function of the number of labels requested ($\log_{10}$ scale). The input space was divided into 10 regions. The curves for ORIWAL are repetitions from Figure 1. The figures show that, given a fixed number of labels, ORIWAL achieves a substantially smaller misclassification loss than IWAL and PASSIVE.

test data against the number of labels requested (on $\log_{10}$ scale), averaged over 50 runs. The error bars indicate ± standard error. ORIWAL shows consistent advantage over RIWAL and RPASSIVE on most datasets, such as magic04, nomao, and ijcnn1, and matches the performance of RIWAL or RPASSIVE on a few others. Since ORIWAL is significantly outperforming the other two region-based algorithms RIWAL and RPASSIVE, for the rest of our experiments we focused on ORIWAL.

We then compared our proposed algorithm ORIWAL with two baselines: the non-region-based IWAL, and the non-region-based passive learning algorithm, PASSIVE. Both IWAL and PASSIVE were run using the hypothesis set $\mathcal{H}$, which is the hypothesis set used in each region of ORIWAL. Figure 2 plots the misclassification error rate achieved by the three algorithms. The optimal region-based algorithm ORIWAL achieves from the beginning a significantly superior prediction accuracy than the two non region-based algorithms, IWAL and PASSIVE. Given the limited space for improvement when working with the single hypothesis set $\mathcal{H}$, IWAL shows no significant improvement over PASSIVE, and stops improving early on. On the other hand, while the learning curve of non region-based algorithms has plateaued, ORIWAL continues to improve in accuracy by leveraging more labels, and manages to significantly outperform PASSIVE and IWAL.

6 Conclusion

We presented a detailed analysis of the scenario of region-based active learning for which we gave a new algorithm, ORIWAL. This algorithm is based on an optimal allocation of points to the underlying region-dependent active learning algorithms. We showed that ORIWAL admits favorable theoretical guarantees, and further demonstrated empirically its substantial improvement over non-region-based algorithms such as IWAL or passive learning in a series of experiments.

Along the way, we also introduced a new active learning algorithm, EIWAL, that benefits from more favorable guarantees than the original IWAL algorithm, and that can be used as a subroutine in our region-based ORIWAL. More generally, other subroutine active learning algorithms can be used with our algorithm, which could lead to further performance improvements in some cases.

We hope to have shown the benefits of region-based active learning and prompted interest in research questions related to this problem. Several crucial questions arise, including the following: How should the regions be chosen? Which hypothesis set should be selected for each? Can we adaptively modify the original partitioning by merging or splitting regions? We have already initiated the study of all of these questions with some preliminary theoretical results. A more complete answer to these and other related questions could lead to significant improvements in active learning.

References

P. Awasthi, M. F. Balcan, and P. M. 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. J. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, pages 199–207, 2010.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees. *CRC Press*, 1984.

N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.

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

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

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

D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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.

T.-K. Huang, A. Agarwal, D. Hsu, J. Langford, and R. E. Schapire. Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems 28*, 2015.

S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

S. Kpotufe, R. Urner, and S. Ben-David. Hierarchical label queries with data-dependent partitions. In *Conference on Learning Theory*, pages 1176–1189, 2015.

J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

P. H. Rossi, J. D. Wright, and A. B. Anderson. *Handbook of Survey Research*. Academic Press, 1983.

R. Urner, S. Wulff, and S. Ben-David. Plal: Cluster-based active learning. In *Conference on Learning Theory*, pages 376–397, 2013.

C. Zhang. Efficient active learning of sparse halfspaces. In *COLT*, 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 of EIWAL

The key step in proving Theorem 1 for EIWAL is using a martingale concentration bound for

$$Z_t = \frac{Q_t}{p_t}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big) - \big(R(f) - R(g)\big),$$

where $Z_1, Z_2, \cdots$ is a martingale difference sequence for any pair $f, g \in \mathtt{H}_T$. Instead of using Azuma's inequality as in [Beygelzimer et al., 2009], we rely on a Berstein-like inequality for martingales Freedman [1975]).

The following result is adapted from Lemma 3 of Kakade and Tewari [2009], which is derived from Freedman [1975]). We denote by $\mathcal{F}_t = \{(x_1, y_1, Q_1), \cdots, (x_t, y_t, Q_t\}$ the observations up to time t.

Lemma 4. *For any $0 < \delta < 1$, and $T \geq 3$, with probability at least $1 - \delta$,*

$$\left|\sum_{t=1}^{T} Z_t\right| \leq \max\left\{2\sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t}[p_t|\mathcal{F}_{t-1}]}, 6\sqrt{\log\left(\frac{8\log(T)}{\delta}\right)}\right\} \times \sqrt{\log\left(\frac{8\log(T)}{\delta}\right)}.$$

Proof. We use Lemma 3 in Kakade and Tewari [2009]. First, observe that variables Z_t are bounded, in particular, $|Z_t| \leq 2$. Furthermore,

$$\begin{aligned}
&\mathrm{var}[Z_t|\mathcal{F}_{t-1}] \\
&= \mathrm{var}\left[\frac{Q_t}{p_t}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big)|\mathcal{F}_{t-1}\right] \\
&\leq \mathop{\mathbb{E}}_{x_t, Q_t}\left[\frac{Q_t^2}{p_t^2}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big)^2|\mathcal{F}_{t-1}\right] \\
&\leq \mathop{\mathbb{E}}_{x_t, Q_t}\left[\frac{Q_t\, p_t^2}{p_t^2}|\mathcal{F}_{t-1}\right] \\
&= \mathop{\mathbb{E}}_{x_t, Q_t}\left[Q_t|\mathcal{F}_{t-1}\right] \\
&= \mathop{\mathbb{E}}_{x_t}\left[p_t|\mathcal{F}_{t-1}\right].
\end{aligned}$$

A union bound over Z_t and $-Z_t$ concludes the proof. □

Given Lemma 4 above, we can adapt Lemma 3 of [Beygelzimer et al., 2009] to using the Berstein-like inequality. Specifically, let us define

$$\Delta_T = \frac{2}{T}\left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right) \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}.$$

Then we have the following high-probability statement for the risk of the hypothesis h_T returned by EIWAL after T rounds.

Lemma 5. *Given any hypothesis class $\mathcal{H}$, for all $\delta > 0$, for all $T \geq 3$ and all $f, g \in \mathtt{H}_T$, with probability at least $1 - 2\delta$,*

$$|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \leq \Delta_T.$$

In particular, if we let $f = h^*$ and $g = h_T$, it follows that

$$R(h_T) \leq R(h^*) + \Delta_T.$$

Proof. Apply Lemma 4 to time $T \geq 3$ and any pair $f, g \in \mathtt{H}_T$, with error probability $\delta/(T^2|\mathcal{H}|^2)$ for round T. A union bound over $T \geq 3$ and (f, g) gives, with probability at least $1 - \delta$,

$$\begin{aligned}
&|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \\
&\leq \frac{1}{T}\max\left\{2\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t}[p_t|\mathcal{F}_{t-1}]}, 6\sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}\right\} \\
&\quad \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}. \qquad (8)
\end{aligned}$$

Next, according to Proposition 2 of Cesa-Bianchi and Gentile [2008], with probability at least $1 - \delta$, for all $T \geq 3$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t}\left[p_t|\mathcal{F}_{t-1}\right] \leq \left(\sum_{t=1}^{T} p_t\right) + 36\log\left(\frac{(3+\sum_{t=1}^{T} p_t)T^2}{\delta}\right) \\
&+ 2\sqrt{\left(\sum_{t=1}^{T} p_t\right)\log\left(\frac{(3+\sum_{t=1}^{T} p_t)T^2)}{\delta}\right)} \\
&\leq \left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right)^2. \qquad (9)
\end{aligned}$$

Combining (8) and (9), we get with probability at least $1 - 2\delta$, for all $T \geq 3$,

$$\begin{aligned}
&|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \\
&\leq \frac{2}{T}\left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right) \\
&\quad \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)},
\end{aligned}$$

as claimed. □

The next lemma gives a label complexity bound for EIWAL.

Lemma 6. *Given any hypothesis class $\mathcal{H}$, and distribution $\mathcal{D}$, with $\theta(\mathcal{D}, \mathcal{H}) = \theta$, for all $\delta > 0$, for all $T \geq 3$, with probably at least $1 - \delta$, we have*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t, Q_t}\left[Q_t|\mathcal{F}_{t-1}\right] \\
&\leq 4\theta K_l\left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right) \\
&\quad + O(\log^3(T|\mathcal{H}|/\delta)),
\end{aligned}$$

where K_ℓ is a constant that depends on ℓ.

Proof. From Theorem 11 of Beygelzimer et al. [2009], for $t \geq 3$,

$$\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq 4\theta K_l (R^* + \Delta_{t-1}), \tag{10}$$

where $R^* = R(h^*)$ is the risk of best-in-class. Plugging in the expression for Δ_{t-1}, and applying again a similar concentration inequality as before to relate $\sum_{t=1}^T p_t$ to $\sum_{t=1}^T \mathbb{E}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right]$, we end up with a recursion on $\mathbb{E}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right]$:

$$\begin{aligned}\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq 4\theta K_l R^* + \frac{4\theta K_l c_1}{t-1}\sqrt{\sum_{s=1}^{t-1}\mathop{\mathbb{E}}_{x_t}\left[p_s \middle| \mathcal{F}_{s-1}\right]} \\ + c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right),\end{aligned} \tag{11}$$

where $c_1 = 2\sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)} = O\left(\sqrt{\log\left(\frac{T|\mathcal{H}|}{\delta}\right)}\right)$, and c_2 is a constant.

For simplicity, denote by $4\theta K_l = c_0$. We show by induction that for all $t \geq 3$,

$$\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq c_0 R^* + c_4\sqrt{\frac{R^*}{t-1}} + \frac{c5}{t-1}, \tag{12}$$

for some constants c_4, c_5. Assume by induction that (12) holds for all $s \leq t-1$. Thus, from (11), we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq c_0 R^* \\ &+ \frac{c_0 c_1}{t-1}\sqrt{c_0 R^*(t-1) + 2c_4\sqrt{R^*(t-1)} + c_5\log(t-1)} \\ &+ c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right) \\ &\leq c_0 R^* + \frac{c_0 c_1}{t-1}\left[\sqrt{c_0 R^*(t-1) + 2c_4\sqrt{R^*(t-1)}}\right. \\ &\left.+ \sqrt{c_5\log(t-1)}\right] + c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right) \\ &\leq c_0 R^* + \frac{c_0 c_1}{t-1}\left[\sqrt{c_0 R^*(t-1)} + \frac{c_4}{\sqrt{c_0}}\right] \\ &+ \frac{c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1} \\ &= c_0 R^* + \frac{c_0 c_1\sqrt{c_0 R^*}}{\sqrt{t-1}} \\ &+ \frac{\sqrt{c_0}c_1 c_4 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1},\end{aligned}$$

where we use the fact that $\sqrt{a+b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$ for $a, b > 0$.

To complete the induction, we need to show that

$$\begin{aligned}&c_0 c_1\sqrt{c_0}\sqrt{\frac{R^*}{t-1}} \\ &+ \frac{\sqrt{c_0}c_1 c_4 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1} \\ &\leq c_4\sqrt{\frac{R^*}{t-1}} + \frac{c5}{t-1}.\end{aligned}$$

Thus, $c_4 = c_0 c_1\sqrt{c_0} = O(\sqrt{\log(T|\mathcal{H}|/\delta)})$, and

$$\begin{aligned}&c_5 \geq c_0^2 c_1^2 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta] \\ &\Rightarrow \sqrt{c_5} = O(c_0 c_1\sqrt{\log T}) \\ &\Rightarrow c_5 = O(c_0^2 c_1^2\log T) = O(\log^2(T|\mathcal{H}|/\delta)).\end{aligned}$$

Thus,

$$\begin{aligned}\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq c_0 R^* + O(\sqrt{\log(T|\mathcal{H}|/\delta)})\sqrt{\frac{R^*}{(t-1)}} \\ + \frac{O(\log^2(T|\mathcal{H}|/\delta))}{t-1}.\end{aligned}$$

Finally,

$$\begin{aligned}&\sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}] = \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t}[p_t|\mathcal{F}_{t-1}] \\ &\leq 4\theta K_l[R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})] \\ &+ O(\log^3(T|\mathcal{H}|/\delta)).\end{aligned} \tag{13}$$

□

Proof of Theorem 1. The bound of generalization error $R(h_T)$ follows from Lemma 5. To get the bound on the number of labels τ_T, we relate $\sum_{t=1}^T \mathbb{E}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}]$ in Lemma 6 to $\tau_T = \sum_{t=1}^T Q_t$ through a Bernstein-like inequality for martingales. Again, from Lemma 3 of Kakade and Tewari [2009], we see that with probability at least $1-\delta$ we have

$$\begin{aligned}&\sum_{t=1}^T Q_t - \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}] \\ &\qquad\leq 2\sqrt{\left(\sum_{t=1}^T \mathrm{var}[Q_t|\mathcal{F}_{t-1}]\right)\log\left(\frac{4\log T}{\delta}\right)} \\ &\qquad\quad + 6\log\left(\frac{4\log T}{\delta}\right) \\ &\qquad\leq 2\sqrt{\left(\sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}]\right)\log\left(\frac{4\log T}{\delta}\right)} \\ &\qquad\quad + 6\log\left(\frac{4\log T}{\delta}\right) \\ &\qquad\leq \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}] + 7\log\left(\frac{4\log T}{\delta}\right).\end{aligned}$$

Combining with (13) completes the proof. □

B Proofs of ORIWAL

Proof of Theorem 2. We first expand the bound in Theorem 1 and get rid of $\sum_{t=1}^{T} p_t$. Lemma 6 states that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t,Q_t}[Q_t|\mathcal{F}_{t-1}] \\ \leq 4\theta K_l \left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right) \\ + O(\log^3(T|\mathcal{H}|/\delta)),$$

which implies that

$$\sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t,Q_t}[Q_t|\mathcal{F}_{t-1}]} \\ \leq \sqrt{4\theta K_l \left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right)} \\ + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)) \\ \leq \sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{1}{2}}(T|\mathcal{H}|/\delta)) \\ + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)) \\ \leq \sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)).$$

Thus, by Theorem 1, with probability at least $1-2\delta$,

$$R(h_T) \\ \leq R(h^*) + \frac{2}{T}\left[\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left[\frac{(3+T)T^2}{\delta}\right]}\right] \\ \times \sqrt{\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\ \leq R(h^*) + \frac{2}{T}\Bigg[\sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)) \\ + 6\sqrt{\log\left[\frac{(3+T)T^2}{\delta}\right]}\Bigg] \times \sqrt{\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\ = R(h^*) + 2\sqrt{\frac{4\theta K_l R(h^*)}{T}\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\ + \frac{O(\log^2(T|\mathcal{H}|/\delta))}{T}.$$

Thus, for each region $\mathcal{X}_k$, with probability at least $1-\frac{\delta}{n}$, for any $T_k > 0$ the following holds:

$$R(h_{k,T}) \\ \leq R_k^* + 2\sqrt{\frac{4\theta_k K_l R_k^*}{T_k}\log\left[\frac{8T_k^2|\mathcal{H}_k|^2\log(T_k)2n}{\delta}\right]} \\ + \frac{O(\log^2(T_k|\mathcal{H}_k|n/\delta))}{T_k}.$$

Recall that

$$R(h_T) = \sum_{k=1}^{n} \mathtt{p}_k R_k(h_{k,T}), \quad R(h^*) = \sum_{k=1}^{n} \mathtt{p}_k R_k^*.$$

A union bound over the n regions gives the result for $R(h_T)$:

$$R(h_T) \leq R(h^*) \\ + \sum_{k=1}^{n} 2\mathtt{p}_k\sqrt{\frac{4\theta_k K_l R_k^*}{T_k}\log\left[\frac{8T_k^2|\mathcal{H}_k|^2\log(T_k)2n}{\delta}\right]} \\ + \Big(\sum_{k=1}^{n}\frac{\mathtt{p}_k}{T_k}\Big)O\Big(\log^2\big(\max_{k\in[n]} T_k|\mathcal{H}_k|n/\delta\big)\Big).$$

Furthermore, from Theorem 1 we have for each region $\mathcal{X}_k$, with probability at least $1-\frac{\delta}{n}$, for any $T_k > 0$:

$$\tau_{k,T} \leq 8\theta_k K_l \left(R_k^* T_k + O(\sqrt{R_k^* T_k \log(T_k|\mathcal{H}_k|n/\delta)})\right) \\ + O\Big(\log^3\big(T_k|\mathcal{H}_k|n/\delta\big)\Big),$$

where $\tau_{k,T}$ denotes the number of labels requested in region k up to time T. Again, a union bound over the n regions gives the result for $\tau_T = \sum_{k=1}^{T}\tau_{k,T}$. □

In order to prove Corollary 3 we need the following standard multiplicative Chernoff bounds.

Theorem 7 (Chernoff)**.** *Let $X_1, \cdots, X_m$ be independent random variables drawn according to some distribution $\mathcal{D}$ with mean $\mathtt{p}$ and support included in $[0,1]$. Then, for any $\gamma \in [0, \frac{1}{p} - 1]$, the following holds for $\widehat{\mathtt{p}} = \frac{1}{m}\sum_{i=1}^{m} X_i$:*

$$\mathbb{P}[\widehat{p} \geq (1+\gamma)p] \leq e^{-\frac{mp\gamma^2}{3}},$$
$$\mathbb{P}[\widehat{p} \leq (1-\gamma)p] \leq e^{-\frac{mp\gamma^2}{2}}.$$

Proof of Corollary 3. Given a total of T queries over all regions, we have $\mathbb{E}[T_k] = T\mathtt{q}_k$, where

$$\mathtt{q}_k = \frac{\mathtt{p}_k\alpha_k}{\sum_{k'}^{n}\mathtt{p}_{k'}\alpha_{k'}}$$

is the probability of querying IWAL$_k$, conditioned on a query being made. By Theorem 7, with probability at least $1-\frac{\delta}{2}$, for all $k \in [n]$,

$$\frac{T_k}{T} \geq \mathtt{q}_k\Big(1 - \sqrt{\frac{2\log(2n/\delta)}{T\mathtt{q}_k}}\Big).$$

It follows that with probability at least $1-\frac{\delta}{2}$, for all $k \in [n]$,

$$\frac{\mathtt{q}_k}{\sqrt{T_k}} = \sqrt{\frac{\mathtt{q}_k}{T}}\sqrt{\frac{\mathtt{q}_k}{(T_k/T)}} \leq \sqrt{\frac{\mathtt{q}_k}{T}}\frac{1}{\sqrt{1-\sqrt{\frac{2\log(2n/\delta)}{T\mathtt{q}_k}}}}.$$

When $T \geq \frac{4\log(2n/\delta)}{\min_{k\in[n]} \mathsf{q}_k}$, we have $\frac{2\log(2n/\delta)}{T\mathsf{q}_k} < \frac{1}{2}$. Since $\frac{1}{\sqrt{1-\sqrt{x}}} \leq 1 + 2\sqrt{x}$ for any $x \leq \frac{1}{2}$, we can write

$$\begin{aligned}\frac{\mathsf{q}_k}{\sqrt{T_k}} &\leq \sqrt{\frac{\mathsf{q}_k}{T}}\Big(1+2\sqrt{\frac{2\log(2n/\delta)}{T\mathsf{q}_k}}\Big)\\ &= \sqrt{\frac{\mathsf{q}_k}{T}} + \frac{2\sqrt{2\log(2n/\delta)}}{T}.\end{aligned}$$

Plugging into Theorem 2, a union bound implies that with probability at least $1-\delta$,

$$\begin{aligned}&R(h_T) \leq R(h^*)\\ &+ 2\sum_{k=1}^{n} \mathsf{p}_k \sqrt{\frac{4\theta_k K_l R_k^*}{T_k} \log\left[\frac{8T^2|\mathcal{H}_k|^2\log(T)4n}{\delta}\right]}\\ &+ \Big(\sum_{k=1}^{n}\frac{\mathsf{p}_k}{T_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T|\mathcal{H}_k|n/\delta\big)\Big)\\ &\leq R(h^*) + 2\sum_{k=1}^{n}\mathsf{p}_k\sqrt{\frac{4\theta_k K_l R_k^*}{T\mathsf{q}_k}\log\left[\frac{8T^2|\mathcal{H}_k|^2\log(T)4n}{\delta}\right]}\\ &+ \Big(\sum_{k=1}^{n}\frac{\mathsf{p}_k}{T\mathsf{q}_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T|\mathcal{H}_k|n/\delta\big)\Big).\end{aligned}$$

Furthermore, by Chernoff bound, with probability at least $1-\delta$, for all $k \in [n]$,

$$\begin{aligned}&T_k \leq T\mathsf{q}_k + \sqrt{3T\mathsf{q}_k\log(n/\delta)}\\ \Rightarrow\ &\sqrt{T_k} \leq \sqrt{T\mathsf{q}_k} + \frac{\sqrt{3T\mathsf{q}_k\log(n/\delta)}}{2\sqrt{T\mathsf{q}_k}}\\ &\quad(\text{ using the inequality } \sqrt{a+b} \leq \sqrt{a} + b/(2\sqrt{a})\ )\\ &\quad\leq \sqrt{T\mathsf{q}_k} + \sqrt{\log(n/\delta)}\,.\end{aligned}$$

Plugging into Theorem 2, with probability at least $1-2\delta$, for any $T > 0$,

$$\begin{aligned}\tau_T \leq& \sum_{k=1}^{n}\Big(8\theta_k K_l\Big[R_k^* T_k + O\Big(\sqrt{R_k^* T_k \log(T_k|\mathcal{H}_k|n/\delta)}\Big)\Big]\\ &+ O\Big(\log^3\big(T_k|\mathcal{H}_k|n/\delta\big)\Big)\Big)\\ \leq& 8K_\ell\Big[\sum_{k=1}^{n}\theta_k R_k^* T\mathsf{q}_k\Big]\\ &+ \sum_{k=1}^{n} O\Big(\sqrt{R_k^* T\mathsf{q}_k \log\Big[\frac{T|\mathcal{H}_k|n}{\delta}\Big]}\Big)\\ &+ O\Big(n\log^3\big(T\max_{k\in[n]}|\mathcal{H}_k|n/\delta\big)\Big).\end{aligned}$$

This concludes the proof. □

C Two Natural Baselines for Region-Based Active Learning

In Section C.1 and Section C.2 below, we analyze two natural extensions of the IWAL algorithm to the region-based setting, called NAIVE-IWAL and RIWAL, that use the composite hypothesis set $\mathcal{H}_{[n]}$ in two different ways. In Section C.3, we then discuss the advantage of RIWAL over NAIVE-IWAL.

The two region-based baselines NAIVE-IWAL and RIWAL can use either IWAL or EIWAL as their underlying subroutines. To avoid clutter in the notation and to simplify the presentation, we proceed with the original version of IWAL, but a similar (though more involved) analysis can be carried out for the enhanced version EIWAL.

C.1 NAIVE-IWAL

NAIVE-IWAL consists of simply running the IWAL algorithm with the composite hypothesis set $\mathcal{H}_{[n]}$. This algorithm will find a model in this set without explicitly taking into account the structure of the set. Despite its simplicity, NAIVE-IWAL admits theoretical guarantees, since the guarantees from the classical IWAL (see Equation (2) and Equation (3)) directly apply. In particular, when $\mathcal{H}_k$s have the same number of hypotheses across k, the complexity terms in these bounds are multiplied by a factor of $\sqrt{n}$. This is because $|\mathcal{H}_{[n]}| = \prod_{k=1}^{n}|\mathcal{H}_k| = |\mathcal{H}_1|^n$. Thus, as the number of regions increases, the complexity term in the bound increases, while the generalization error of the best in class $R(h^*)$ decreases.

C.2 RIWAL

RIWAL consists of running n separate IWAL algorithms independently for each region. It works exactly in the same way as ORIWAL, except that it simply passes on all points to the subroutines, that is $\alpha_k = 1$ for all $k \in [n]$. Given T_k, which is the number of samples falling into region $\mathcal{X}_k$, RIWAL admits the same generalization error guarantees as that of ORIWAL (Theorem 2). Both results are derived from IWAL for a single region, along with a union bound over n regions. We can also apply a multiplicative Chernoff bound to the empirical quantities T_k to obtain a learning guarantee that only depends on T. The result is in fact a special case of Corollary 3, and is obtained by simply replacing therein q_k with p_k.

C.3 Comparing NAIVE-IWAL and RIWAL

Even though NAIVE-IWAL and RIWAL learn from the same hypothesis set $\mathcal{H}_{[n]}$, and essentially use the same policy (the disagreement-based policy of IWAL) for requesting labels, the two algorithms are not equivalent. In fact, the two algorithms deliver final hypotheses with comparable generalization error after T rounds but, as we will show momentarily, NAIVE-IWAL request more labels than RIWAL in expectation.

The following definitions will be useful. Let $\widehat{R}_{k,t}(h)$ and $\widehat{R}_t(h)$ denote the importance weighted empirical error of any hypothesis h after t rounds on region $\mathcal{X}_k$ and over all regions, respectively:

$$\widehat{R}_{k,t}(h) = \frac{\sum_{s=1}^t 1_{x_s \in \mathcal{X}_k} \frac{Q_s}{p_s} \ell(h(x_s), y_s)}{\sum_{s=1}^t 1_{x_s \in \mathcal{X}_k}},$$

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

Let $\widehat{h}_{k,t}$ and $\widehat{h}_t$ be the respective weighted empirical risk minimizers:

$$\widehat{h}_{k,t} = \operatorname*{argmin}_{h \in \mathcal{H}_k} \widehat{R}_{k,t}(h), \;\; \widehat{h}_t = \operatorname*{argmin}_{h \in \mathcal{H}_{[n]}} \widehat{R}_t(h) .$$

Similar to Equation (1), we have $\widehat{h}_t = \sum_{k=1}^n 1_{x \in \mathcal{X}_k} \widehat{h}_{k,t}$.

Recall that for NAIVE-IWAL and RIWAL, the probability of requesting label y_t depends on the "disagreement" among their version spaces on x_t. A larger version space implies a larger disagreement value, and therefore a larger probability of requesting the label. Thus, at a high level, NAIVE-IWAL requests more labels than RIWAL because the version space of NAIVE-IWAL is larger than that of RIWAL. More precisely, assume for now that NAIVE-IWAL and RIWAL have been requesting the same labels up to time $t-1$, thus for any h and k, $\widehat{R}_{k,t}(h)$ has the same value under either algorithm, and the region-specific empirical risk minimizer is $\widehat{h}_{k,t}$. At time t, assume without loss of generality, that the unlabeled x_t lies in region $\mathcal{X}_1$. Given a slack term Δ, the version space is defined as the set of hypotheses whose importance weighted empirical error is Δ-close to the minimal empirical error. Assume there exists a hypothesis $h_1 \in \mathcal{H}_1$ such that

$$\Delta \le \widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t}) \le \left[\frac{t}{\sum_{s=1}^t 1_{x_s \in \mathcal{X}_1}}\right] \Delta.$$

Since $\Delta \le \widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t})$, h_1 will not be included in the current version space of IWAL$_1$, which is the subroutine associated with $\mathcal{X}_1$ under the RIWAL algorithm. However, the version space of NAIVE-IWAL will include the hypothesis that takes the value of h_1 on region $\mathcal{X}_1$. To see why, let

$$h' = \sum_{k \in [n], k \neq 1} 1_{x \in \mathcal{X}_k} \widehat{h}_{kt} + 1_{x \in \mathcal{X}_1} h_1 ,$$

that is, the hypothesis that takes the value of the region-specific weighted empirical risk minimizers ($\widehat{h}_{k,t}$) on region $\mathcal{X}_k$, and takes the value of h_1 on region $\mathcal{X}_1$. Since

$$\begin{aligned} &\widehat{R}(h') - \widehat{R}(\widehat{h}_t) \\ &= \left[\frac{\sum_{s=1}^t 1_{x_s \in \mathcal{X}_1}}{t}\right] \left(\widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t})\right) \le \Delta, \end{aligned}$$

h' will be included in the version space of NAIVE-IWAL under the slack term Δ, even though h_1 is not included in the version space of RIWAL on region $\mathcal{X}_1$ under the same slack term. This suggests that NAIVE-IWAL is less efficient at shrinking the version space, and as a result it requests more labels.

We formalize this idea with Lemma 8 and Theorem 9. Lemma 8 relates the region-specific disagreement coefficients $\theta(\mathcal{D}_k, \mathcal{H}_k)$ to the overall disagreement coefficient $\theta(\mathcal{D}, \mathcal{H}_{[n]})$. Theorem 9 compares the learning guarantees of NAIVE-IWAL and RIWAL under certain assumptions.

Lemma 8. *The generalized disagreement coefficient $\theta(\mathcal{D}, \mathcal{H}_{[n]})$ satisfies $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \le \sum_{k=1}^n \theta(\mathcal{D}_k, \mathcal{H}_k)$.*

Proof. Denote $h^* = \operatorname{argmin}_{h \in \mathcal{H}_{[n]}} R(h)$, and $h_k^* = \operatorname{argmin}_{h \in \mathcal{H}_k} R_k(h)$. For simplicity, we denote by $\mathcal{D}_k = \mathcal{D}|\mathcal{X}_k$ the conditional distribution of x on $\mathcal{X}_k$. Recall that $h^* = \sum_{k=1}^n 1_{x \in X_k} h_k^*$. Extending the definitions in Section 3, we define

$$\rho_k(f, g) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \max_y |\ell(f(x), y) - \ell(g(x), y)|.$$

Given the hypothesis set $\mathcal{H}_k$ and any real $r > 0$, define

$$B_k(f, r) = \big\{g \in \mathcal{H}_k \colon \rho_k(f, g) \le r\big\}.$$

For a set of non-negative values $\lambda = \{\lambda_1, \ldots, \lambda_n\}$, let

$$G_\lambda(h^*, r) = \Big\{\sum_{k=1}^n 1_{x \in X_k} g_k \colon g_k \in B_k(h_k^*, \lambda_k r)\Big\}.$$

We first show that, for any λ satisfying $\sum_{k=1}^n \mathrm{p}_k \lambda_k \le 1$, $G_\lambda(h^*, r) \subseteq B(h^*, r)$. Let $g = \sum_{k=1}^n 1_{x \in X_k} g_k$, where $g_k \in B_k(h_k^*, \lambda_k r)$. Then,

$$\begin{aligned} &\rho\left(h^*, g\right) \\ &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \max_y |\ell(h^*(x), y) - \ell(g(x), y)| \\ &= \sum_{k=1}^n \mathrm{p}_k \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \max_y |\ell(h_k^*(x), y) - \ell(g_k(x), y)| \\ &\le \sum_{k=1}^n \mathrm{p}_k \lambda_k r \le r. \end{aligned}$$

Thus, $\left\{ \cup_{\lambda \colon \sum_{k=1}^n \mathrm{p}_k \lambda_k \le 1} G_\lambda(h^*, r) \right\} \subseteq B(h^*, r)$. On the other hand, if there exits a hypothesis h such that

$$h \in B(h^*, r) \Big\backslash \left\{ \cup_{\lambda \colon \sum_{k=1}^n \mathrm{p}_k \lambda_k \le 1} G_\lambda(h^*, r) \right\},$$

let $h = \sum_{k=1}^n 1_{x \in X_k} h_k$. Then,

$$\rho(h^*, h) = \sum_{k=1}^n \mathrm{p}_k \rho_k(h_k^*, h_k) \le r \Rightarrow \sum_{k=1}^n \mathrm{p}_k \frac{\rho_k(h_k^*, h_k)}{r} \le 1.$$

Obviously, $h_k \in B_k(h_k^*, \rho_k(h_k^*, h_k))$. Thus, let $\lambda = \{\frac{\rho_1(h_1^*,h_1)}{r}, \ldots, \frac{\rho_p(h_n^*,h_n)}{r}\}$, then $\sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1$, and $h \in G_\lambda(h^*, r)$ by definition. We have a contradiction. Therefore,

$$\left\{ \cup_{\lambda\colon \sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r) \right\} = B(h^*, r) \, .$$

Given the equivalence above, for any $k \in [n]$,

$$\begin{aligned} \mathcal{H}_k \cap B(h^*, r) &= \mathcal{H}_k \cap \{\cup_{\lambda\colon \sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r)\} \\ &= \mathcal{H}_k \cap \{\cup_{\lambda_k \leq 1/\mathtt{p}_k} B_k(h_k^*, \lambda_k r)\} \qquad (14) \\ &= B_k(h_k^*, r/\mathtt{p}_k) \, . \qquad (15) \end{aligned}$$

Equation (14) holds by the definition of $G_\lambda(h^*, r)$. Putting everything together, we have for any $r \geq 0$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{x \sim D} \sup_{h \in B(h^*, r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \\ &= \sum_{k=1}^n \mathtt{p}_k \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \sup_{h \in B(h^*, r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \\ &= \sum_{k=1}^n \mathtt{p}_k \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \sup_{y, h_k \in B_k(h_k^*, \frac{r}{\mathtt{p}_k})} |\ell(h_k(x), y) - \ell(h_k^*(x), y)| \\ &\qquad (16) \\ &\leq \sum_{k=1}^n \mathtt{p}_k \theta(\mathcal{D}_k, \mathcal{H}_k) r/\mathtt{p}_k \qquad (17) \\ &= \Big(\sum_{k=1}^n \theta(\mathcal{D}|\mathcal{X}_k, \mathcal{H}_k) \Big) r. \end{aligned}$$

Equation (16) holds due to the equivalence in (15), and inequality (17) holds by the definition of $\theta(\mathcal{D}_k, \mathcal{H}_k)$.

Finally, recall the definition of $\theta(\mathcal{D}, \mathcal{H}_{[n]})$:

$$\begin{aligned} \theta(\mathcal{D}, \mathcal{H}) = \inf \Big\{ \theta \colon \forall r \geq 0, \\ \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \sup_{h \in B(h^*, r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \leq \theta r \Big\}. \end{aligned}$$

Therefore $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \leq \sum_{k=1}^n \theta(\mathcal{D}_k, \mathcal{H}_k)$, which conclues the proof. □

In fact, one can show that there exist r, $\mathcal{D}$ and $\mathcal{H}_k$ such that equality is achieved in Lemma 8, thus the upper bound is tight.

Combining Lemma 8 with the learning guarantee of IWAL, we obtain the following result for the case when $|\mathcal{H}_k|$ is the same across all regions $\mathcal{X}_k$.

Theorem 9. *Assume $|\mathcal{H}_k|$ is the same across all regions $\mathcal{X}_k, k \in [n]$, and assume the same holds for $\theta(\mathcal{D}_k, \mathcal{H}_k)$. Then, the hypothesis returned by* NAIVE-IWAL *and* RIWAL *admit comparable generalization error guarantees, but on average* NAIVE-IWAL *would request up to n times more labels than* RIWAL.

Proof. Let $N = |\mathcal{H}_1|$, and $\theta_1 = \theta(\mathcal{D}_1, \mathcal{H}_1)$, so that $|\mathcal{H}_{[n]}| = N^n$ and, from Lemma 8, $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \leq n\theta_1$. According to the learning guarantee of IWAL, with probability at least $1 - \delta$, NAIVE-IWAL satisfies

$$R(h_T^{\text{NAIVE-IWAL}}) \leq R(h^*) + O\Big(\sqrt{\frac{\ln(TN^{2n}/\delta)}{T}}\Big), \qquad (18)$$

$$\tau_T^{\text{NAIVE-IWAL}} \leq 4n\theta_1 K_\ell \Big[R(h^*) T + O(\sqrt{T \ln(TN^{2n}/\delta)}) \Big]. \qquad (19)$$

Meanwhile according to Theorem 2, with probability at least $1 - \delta$, RIWAL satisfies

$$\begin{aligned} &R(h_T^{\text{RIWAL}}) \\ &\quad \leq R(h^*) + \sum_{k=1}^n \mathtt{p}_k O\Big(\sqrt{\frac{\ln(T|N|^2 n/\delta)}{T_k}}\Big), \qquad (20) \\ &\tau_T^{\text{RIWAL}} \\ &\quad \leq \sum_{k=1}^n 4\theta_1 K_\ell \Big[R_k(h^*) T \mathtt{p}_k + O(\sqrt{2T\mathtt{p}_k \ln(2TN^2 n/\delta)} \Big] \\ &\quad = 4\theta_1 K_\ell \Big[R(h^*) T + \sum_{k=1}^n O(\sqrt{2T\mathtt{p}_k \ln(2TN^2 n/\delta)}) \Big]. \\ &\qquad (21) \end{aligned}$$

Replacing T_k with $T\mathtt{p}_k + O(\sqrt{T})$ in the RHS of (20) we obtain

$$R(h_T^{\text{RIWAL}}) \leq R(h^*) + O\Big(\sqrt{\frac{n \ln(T|N|^2 n/\delta)}{T}}\Big). \qquad (22)$$

Comparing the upper bound on the generalization error of RIWAL (22) to that of NAIVE-IWAL (18), we conclude that the two algorithms admit comparable learning guarantees.

On the other hand, comparing the proportion of labels requested per round, we have

$$\begin{aligned} \tau_T^{\text{NAIVE-IWAL}}/T &\leq 4n\theta_1 K_\ell R(h^*) + O\Big(\frac{1}{\sqrt{T}}\Big), \\ \tau_T^{\text{RIWAL}}/T &\leq 4\theta_1 K_\ell R(h^*) + O\Big(\frac{1}{\sqrt{T}}\Big). \end{aligned}$$

Thus, NAIVE-IWAL may request up to n times more labels than RIWAL.

□

D More Experimental Results

In this section, we provide results for all the datasets described in Table 1 in the main body of the paper.

Figures 3 show for 10 disjoint regions the misclassification error rate by three region-based algorithms, ORIWAL, RIWAL, and RPASSIVE, against number of labels requested (on

$\log_{10}$ scale), for all datasets. ORIWAL displays a consistent advantage over RIWAL and RPASSIVE.

Figures 4 and Figure 5 compares, for 10 and 20 disjoint regions respectively, the misclassification error rate of our algorithm, ORIWAL, to that of non region-based IWAL, and to non region-based passive learning PASSIVE. IWAL performs comparably to PASSIVE and stops improving early on, while ORIWAL significantly outperforms PASSIVE and continues to reduce the error rate while requesting more labels.

Figures 6 show the misclassification error rate by ORIWAL using 10 regions and 20 regions, respectively, against number of labels requested (on $\log_{10}$ scale), for all datasets. With randomly generated regions, it is unclear whether more regions would be helpful, as sometimes 20 regions admit higher misclassification error compared to 10 regions, given the same amount of requested labels. This observation leads to the following questions: How should the regions be chosen? How would the partitioning method affect the performance of ORIWAL? These are interesting directions for future work.

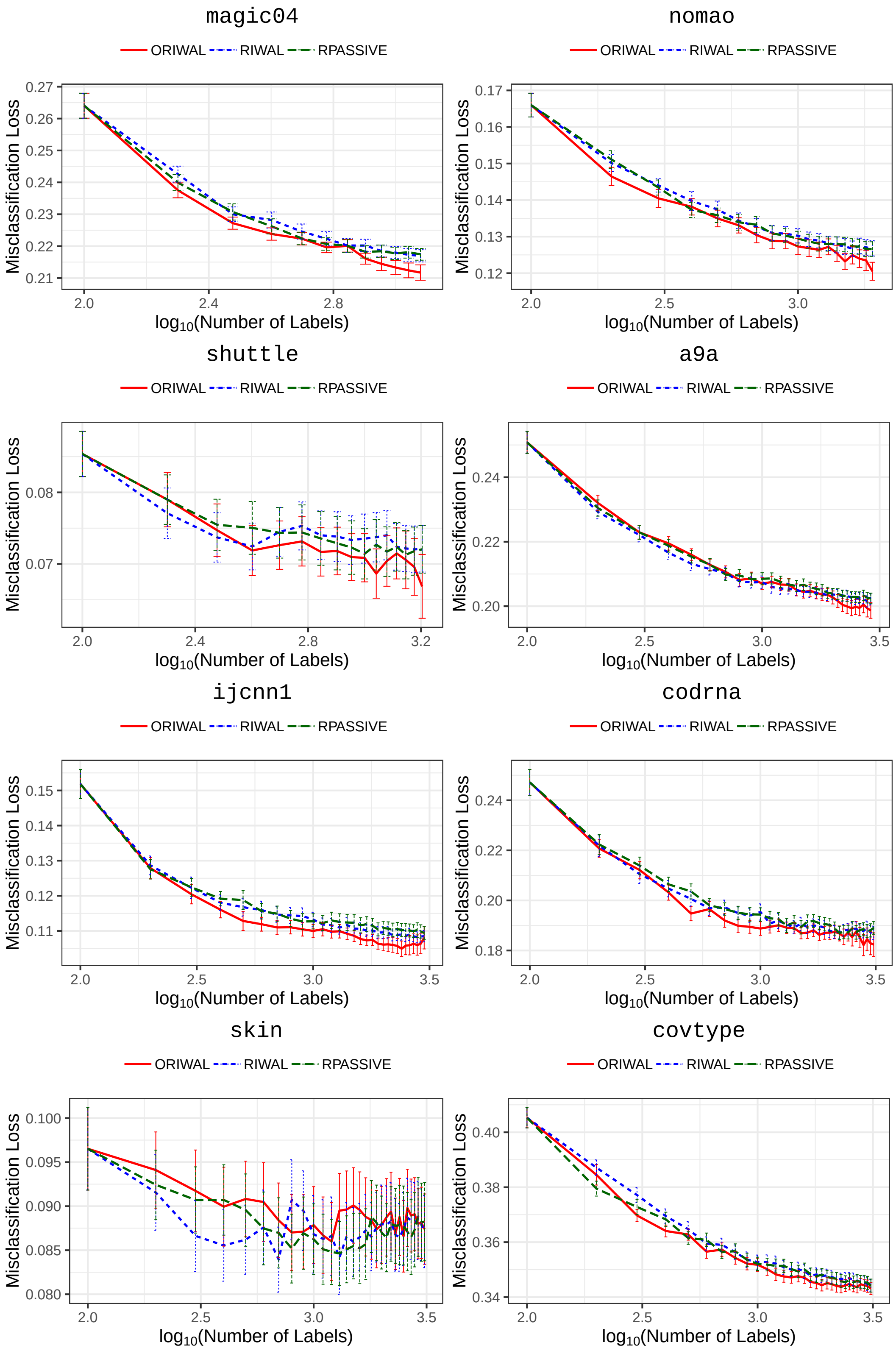

Figure 3: Misclassification loss of ORIWAL, RIWAL, and RPASSIVE on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **10** regions.

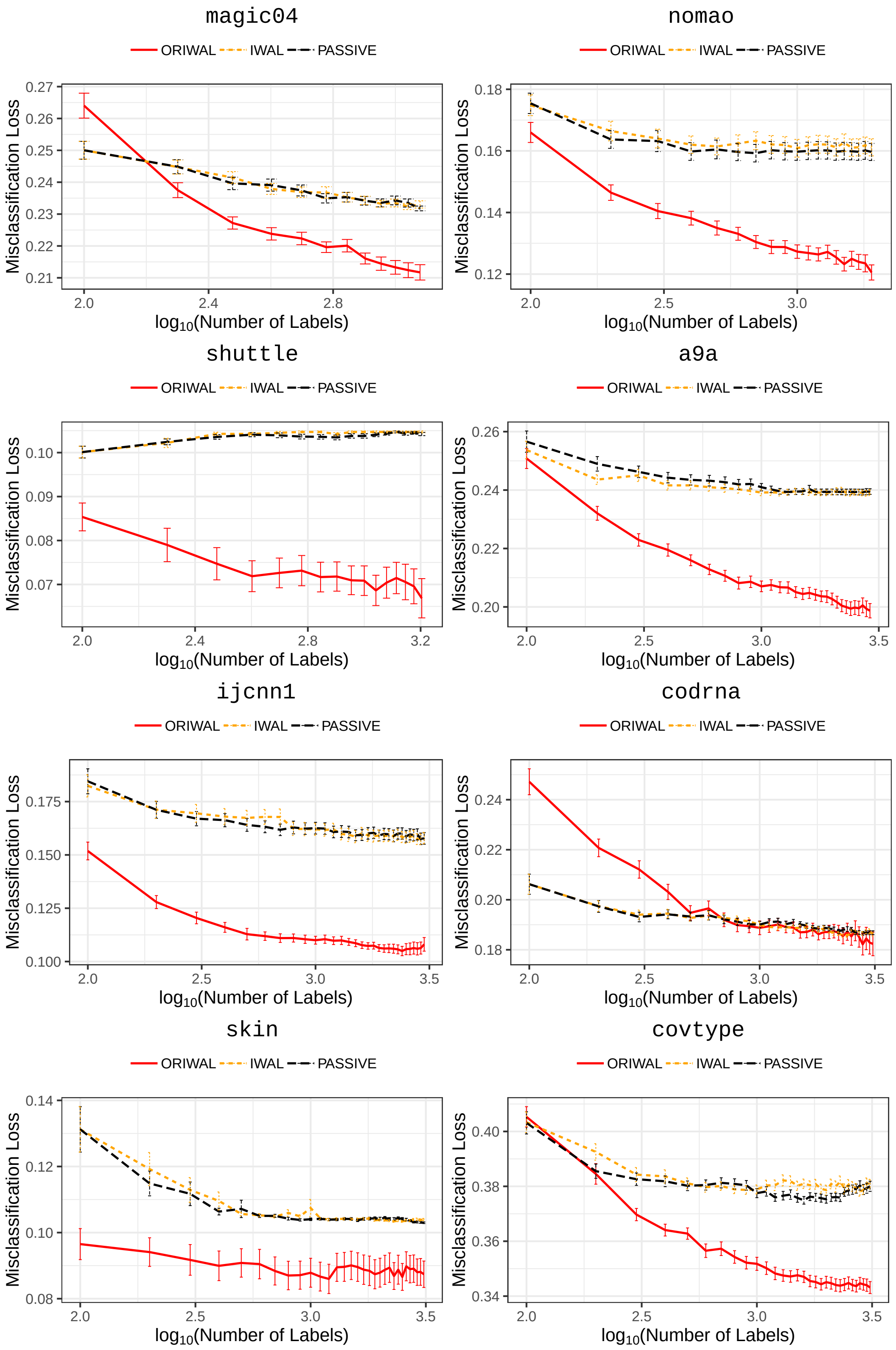

Figure 4: Misclassification loss of non region-based IWAL, non region-based passive learning PASSIVE, and ORIWAL (ours) on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **10** regions.

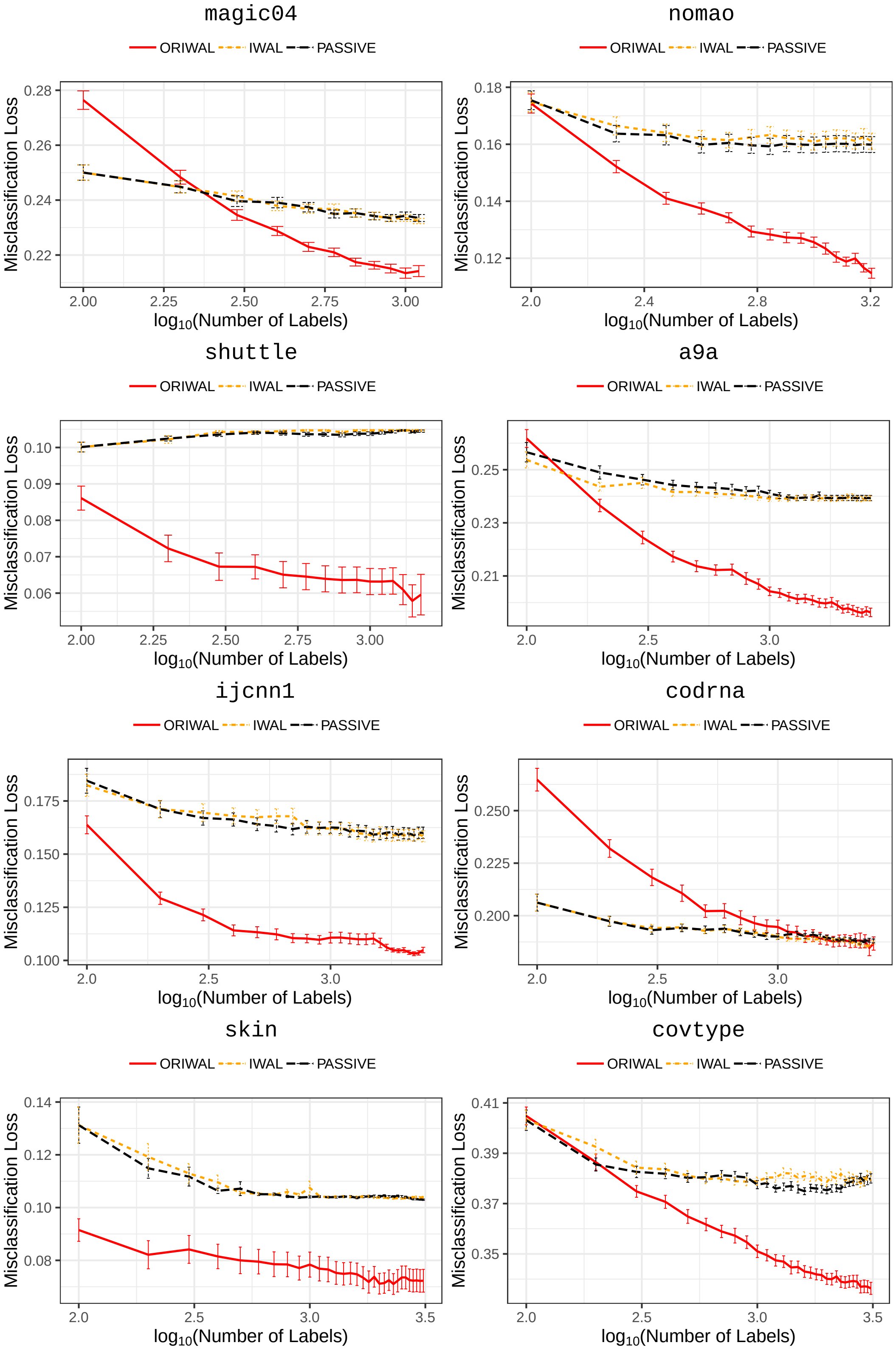

Figure 5: Misclassification loss of non region-based IWAL, non region-based passive learning PASSIVE, and ORIWAL (ours) on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **20** regions.

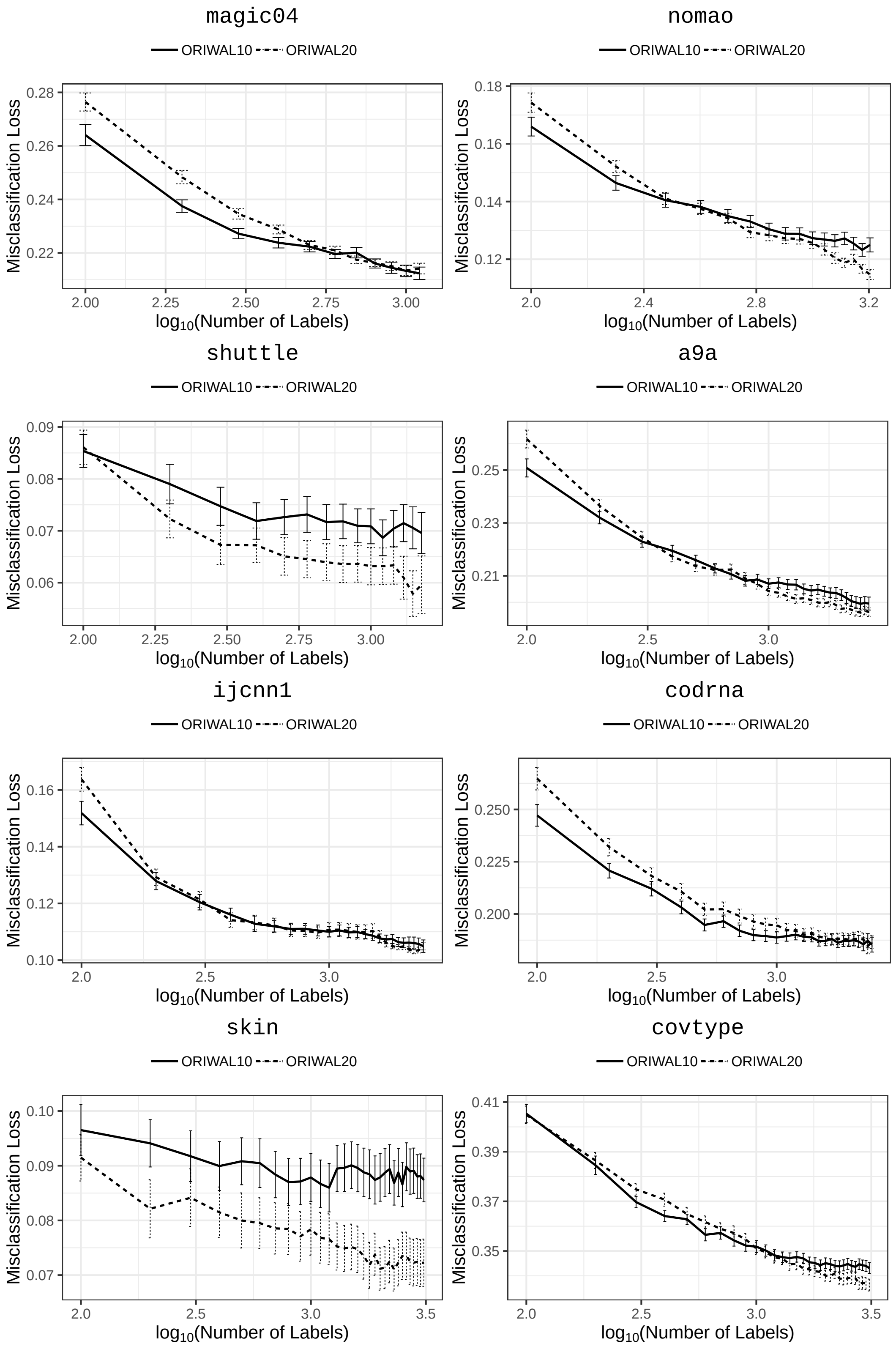

Figure 6: Misclassification loss of ORIWAL, using 10 regions, vs. 20 regions, on hold out test data vs. number of labels requested ($\log_{10}$ scale).