Active Learning under the Bernstein Condition for General Losses

by

Hamid Shayestehmanesh
B.Sc., Shiraz University, 2017

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Active Learning under the Bernstein Condition for General Losses

by

Hamid Shayestehmanesh
B.Sc., Shiraz University, 2017

Supervisory Committee

Dr. Nishant Mehta, Supervisor
(Department of Computer Science)

Dr. Kwang Moo Yi, Departmental Member
(Department of Computer Science)

ABSTRACT

We study online active learning under the Bernstein condition for bounded general losses and offer a solution for online variance estimation. Our suggested algorithm is based on IWAL (Importance Weighted Active Learning) which utilizes the online variance estimation technique to shrink the hypothesis set. For our algorithm, we provide a fallback guarantee and prove that in the case that $R(f^*)$ is small, it will converge faster than passive learning, where $R(f^*)$ is the risk of the best hypothesis in the hypothesis class. Finally, in the special case of zero-one loss exponential improvement is achieved in label complexity over passive learning.

Contents

Supervisory Committee ii

Abstract iii

Table of Contents iv

List of Figures vi

Acknowledgements vii

1 Introduction 1

2 Background 4

3 IWAL-σ 8

3.1 Introduction . . . 8

3.2 Preliminaries . . . 9

3.3 Review of IWAL . . . 9

3.4 IWAL-σ . . . 11

3.5 Label Complexity . . . 22

3.6 Special case of zero-one loss . . . 27

3.7 Future Work . . . 34

4 Experiments 35

4.1 Experiment Setup . . . 35

4.2 Results and Discussion . . . 38

4.2.1 Zero-one Loss . . . 38

4.2.2 Hinge loss . . . 40

4.2.3 Logistic and Squared Loss . . . 43

4.3 Conclusion . . . 47

5 Conclusion and Future Work 48

A Additional Information 49

A.1 Lemmas 49

Bibliography 50

List of Figures

Figure 3.1 Dependency Graph 19

Figure 4.1 Results for zero-one loss with no modification on IWAL-σ . . . 38
Figure 4.2 Results for zero-one loss for modified IWAL-σ 39
Figure 4.3 Excess risk of the worst function left in version space for modified IWAL-σ 40
Figure 4.4 Results for hinge loss with no modification on IWAL-σ 41
Figure 4.5 Results for hinge loss for modified IWAL-σ - 5% noise 41
Figure 4.6 Excess risk of the worst function left in the effective version space for modified version of IWAL-σ - Hinge loss - 5% noise 42
Figure 4.7 Results for logistic loss with no modification on IWAL-σ 43
Figure 4.8 Results for squared loss with no modification on IWAL-σ 44
Figure 4.9 Results for the modified version of IWAL-σ 44
Figure 4.10Excess risk of the worst function left in the effective version space for modified IWAL-σ - Logistic loss - 5% noise 45
Figure 4.11Excess risk of the worst function left in the effective version space for the modified version of IWAL-σ - Squared loss - 5% noise . 46

ACKNOWLEDGEMENTS

I would like to thank all the committee members who reviewed my work as well as my family, especially my brother, who supported me full-heartedly throughout my degree. Dr. Kwang Moo Yi provided useful feedback on my thesis; I appreciate his remarks and wish him well. A very special thanks to Dr. Steve Hanneke, who graciously accepted to act as my examiner for my defense and went above and beyond with his comments and critiques. I am very thankful that he did not hold back and helped me produce the best work I could have during this process. Last but not least, I would like to personally thank my supervisor, Dr. Nishant Mehta; his continued support, validation, and drive have pushed me to grow and achieve success throughout my degree.

Hamid Shayestehmanesh

Chapter 1

Introduction

A good dataset is essential for solving a learning problem. Many supervised learning algorithms, such as neural networks, require not only a labeled dataset but also a large one. However, in some situations, we have access to cheap unlabeled data, but labeling the data is expensive, time-consuming or difficult for some other reasons. For example, assume the problem of annotating semi-scientific texts of a particular subject on Wikipedia. One needs to hire an expert in the related majors for annotating the data. Clearly, the cost of hiring experts would be high. Another example is learning problems related to medical issues. Collecting data from patients is difficult and can take up to years; therefore, one would like to make sure these samples are as informative as possible. Usually, an analyst would like to solve their learning problem as cheap and as fast as possible. A solution in such situations is to use unsupervised learning. However, unsupervised learning algorithms have shown promising results in specific problems; it is not possible to utilize them in many cases. In many cases, unsupervised learning algorithms are usually built upon prior knowledge about the data or the problem, which, if this knowledge is unavailable, then unsupervised learning algorithms are not guaranteed to do well. Thus, we use active learning to label as few samples as possible while utilizing the unlabeled samples.

Active learning is a subfield of semi-supervised learning. In semi-supervised learning, the learning is provided with labeled and unlabeled samples. It is reasonable to assume the labeled samples are more limited than the unlabeled samples. In many semi-supervised learning problems, the labeled and unlabeled data is simply given to the algorithm by Nature. In active learning, however, the algorithm can interactively choose what unlabeled samples to label. The most crucial active learning algorithm's goal is to label as few samples as possible and speed up the learning process. To do so, the algorithm must query the most informative samples. Informative samples are the most useful samples for the algorithm, which depends on how the algorithm works. Some of these techniques will be discussed later in Chapter 3.

Assume the following realizable classification problem. The samples are on a 1-d space. For any sample $x \leq C$, it is labeled 0; otherwise, it is labeled 1. We aim to find the best 1-d threshold function. It is shown that an active learning algorithm can find a threshold function with low excess risk by only querying $\log(n)$ labels, where n is the number of samples required by passive learning to find a threshold function as good (Dasgupta, 2005). The idea for this algorithm is similar to the binary search algorithm. However, it might not be possible to expand this idea to other learning problems, but it shows how active learning can decrease the number of necessary labels to solve a problem. Some of the most popular fields that have utilized active learning is natural language processing (NLP), Computer Vision, and Biomedical Imaging. Many NLP applications, such as Part-of-Speech Tagging or Named Entity Recognition, require many labeled samples. More importantly, labeling these samples are expensive as it is time-consuming, and only an expert can label the samples.

Active learning has been studied in many different settings. In this work, we first study active learning under the Bernstein noise condition for general losses. Learning under the Bernstein noise condition is well studied in passive learning, and it has been shown that under the Bernstein noise condition, we can achieve better label complexities since the noise is limited. We discuss under what conditions, our proposed algorithm can achieve a label complexity smaller than passive learning. A similar study has been done for zero-one loss in active learning framework (Huang et al., 2015); in the last section of chapter 3, we recover their results by slightly modifying our algorithm. Second, we introduce an algorithm that improves upon a fundamental algorithm in active learning community, IWAL (Beygelzimer et al., 2009). Our second algorithm is aimed for harder problems, while the first algorithm is intended to be used for data with bounded noise.

Outline. In **Chapter 2**, we briefly discuss different settings of active learning and some of the important works around each topic. In **Chapter 3** is the main contribution of this thesis. We propose an algorithm designed to benefit from the Bernstein condition for bounded general losses. We also modify our algorithm for zero-one loss to achieve up to exponential improvements in label complexity over passive learning, under the Bernstein condition. **Chapter 4** consists of experiments to compare our algorithm to passive learning and also, to study the behavior of our algorithm in different cases. Finally, in **Chapter 5** , we have discussion and possible future work.

Chapter 2

Background

An active learning algorithm interactively chooses whether to query the label of an unlabeled sample from an oracle. A good active learner must query as few labels as possible to learn. This technique is especially applicable to problems where an enormous amount of unlabeled data exists, but labeling the data is expensive, time-consuming, or, overall, difficult to label by a realistic oracle. Such problems are often observed in natural language processing, computer vision, and many other important areas.

Active learning has been studied in different settings. Two of the most important settings are the pool-based and online (sequential) ones, where the main difference is how the learner observes samples. In the pool-based setting, a learner has access to all the unlabeled data at once. At every iteration, the learner picks one or a small batch of samples from the sample set and asks the oracle to label them. It could be said that pool-based active learning is inspired by the batch setting in passive learning. Similarly, online active learning is inspired by online learning. In the online setting, which is the focus of this study, the learner observes a sample at each round and right away must decide whether to query the label or not. If the learner skips querying that sample, she will never be able to request the oracle to label it again. Two kinds of guarantees should be given for active learning algorithms: (i) an upper bound on the generalization error of the returned hypothesis $\hat{f}$ by the algorithm; (ii) an upper bound on label complexity, i.e., the number of labels required to achieve generalization error at most ϵ. Here the generalization error is the difference in the error of hypothesis returned by the algorithm and the best hypothesis in the hypothesis set.

Another way to categorize active learning works is by their approach to the problem. Three main ideas have been used by the active learning community: *disagreement-*

based, *margin-based*, and *cluster based*. The margin-based line of work is mostly dedicated to linear separators. For a short survey on these techniques, see (Balcan and Urner, 2016).

Different approaches have been taken to tackle the pool-based setting. One of the most significant techniques is *exploiting cluster structure in the data*. The general idea is to separate the data into m clusters, where each cluster represents only one class. Note that more than one cluster can represent a class. A good algorithm based on this technique should be able to, first, break the data into adequately small clusters so that each cluster is almost pure without making them too small. By pure, we mean that most of the samples in each cluster are from one class. Second, the algorithm should not need to know m in advance. Dasgupta and Hsu (2008) have proposed an algorithm that can find such clusters. Their algorithm queries a batch of samples at each round of learning, and then, in a bottom-up fashion, the algorithm scores each cluster, and if necessary, the algorithm prunes the cluster to two smaller clusters. This process is repeated until no cluster needs to be pruned, or the algorithm has exhausted the labeling budget; where the labeling budget is the number of labels an algorithm is allowed to query from the oracle. They show if there exists a clustering of m clusters with error η, then the algorithm is guaranteed to find it after $O(m/\eta)$ label queries. Once the algorithm is finished, we can label each cluster by the label associated with it. Finally, this labeled dataset could be used by any passive learning algorithm that is robust to errors in a small dataset.

Another technique used to approach the pool-based setting is the disagreement technique. This technique is used in both frameworks, online active learning and pool-based active learning. Our contributions, discussed in Chapter 3, are based on the disagreement technique and concern the online active learning setting. For this reason, we discuss some of the disagreement-based works in online active learning in more detail in Chapter 3. Most algorithms based on this technique, implicitly or explicitly, maintain a version space. The version space is a subset of the hypothesis set such that the algorithm believes any hypothesis in the version space has excess risk lower than a certain amount, where the excess risk of a hypothesis f is the difference of risk of the best hypothesis in the hypothesis set and the risk of f (Generalization error, defined in the previous paragraph, is the excess risk of the returned hypothesis by the algorithm). As the algorithm visits more labeled data, it intends to shrink the version space. There are two common challenges in this line of work. First, algorithms that maintain the version space explicitly are computationally intractable

(Beygelzimer et al., 2009; Dasgupta, 2006). Second, these algorithms tend to be mellow in their decisions, i.e, they query almost any sample if there are two hypotheses in the version space disagreeing on its label (Beygelzimer et al., 2009, 2010; Dasgupta, 2006). Beygelzimer et al. (2010); Huang et al. (2015) have proposed algorithms that avoid explicit use of a version space and that are computationally efficient considering an ERM oracle. However, we suspect that it is possible to implement most of these algorithm in a way to avoid explicit implementation of version space.

Tosh and Dasgupta (2017); Dasgupta (2006) have used the disagreement technique in the pool-based setting. The work of Tosh and Dasgupta (2017) is of interest because not only do they manage to propose an efficient algorithm, but their procedure for querying a label is more aggressive than similar works, and most importantly, their idea could inspire a faster algorithm in online active learning. In this subject, we refer to an algorithm as aggressive, if this algorithm has a stingier policy for querying labels. For example, if an algorithm would query a label only based on the outcome of two hypotheses in the version space, this algorithm is mellow; however, if an algorithm considers the outcome of all the hypotheses to decided whether it should query a label, then most likely this algorithm will end up querying fewer labels and is more aggressive. At every round, Tosh and Dasgupta (2017)'s algorithm asks the oracle to label the sample with the maximum average disagreement among the hypotheses left in the version space.

The disagreement based technique has been used in the online active learning setting for many years (Balcan et al., 2006; Beygelzimer et al., 2010, 2009; Huang et al., 2015; Cortes et al., 2019b,a). Balcan et al. (2006) proposed an algorithm called A^2, which was the first work to propose an algorithm that its only assumption was that the samples are i.i.d instead of older works with stronger assumptions like realizability. Later, Hanneke (2007) showed an upper bound on label complexity of A^2. A common phenomenon in these works is that the number of labels required to achieve generalization error ϵ consists of at least two terms, first, a term that depends on the generalization error ϵ. Second, a term that depends on $R(f^*)$, the risk of the best hypothesis in the hypothesis set. To our knowledge, a general efficient algorithm with good generalization bounds does not exist.

An important question to ask in active learning is *"when can active learning help?"* One way to answer this question is by studying the label complexity of a problem and the importance of unlabeled data. Dasgupta (2006) studied the sample complexity of an active learning problem which depends on (1) the desired accuracy ϵ; (2) the distribution over the input space; and (3) the target function f^* in the hypothesis set $\mathcal{H}$ for problems with finite VC dimension or separable problems. The notion of *splitting index* (ρ, ϵ, τ), introduced in this work, is meant to capture the local geometry of $\mathcal{H}$ around the f^*. They prove that the sample complexity of an active learning problem is $\Omega(1/\rho)$. More interestingly, they show, for a hypothesis set $\mathcal{H}$ with splitting index (ρ, ϵ, τ), to learn a hypothesis with an error at most ϵ, any active learning algorithm with $\leq 1/\tau$ unlabeled samples needs to request at least $1/\rho$ labels. These results show us that a lack of unlabeled samples increases the required amount of labeled samples, but it does not show that the existence of unlabeled data can reduce the required amount of labeled samples. This question is also well studied in passive learning (Göpfert et al., 2019; Kääriäinen, 2005).

Active learning has been studied beyond binary classification. Agarwal (2013) studies the problem of cost-sensitive multi-class classification. Multi-class classification in the context of active learning was mostly studied with an empirical approach before this work. Agarwal also looks at multi-class learning under a low noise condition. Another area studied in active learning is the region-based setting. In this setting, the input space is partitioned into regions. Cortes et al. (2019a) studies this setting and proposes an algorithm based on IWAL, which learns a different hypothesis for each region.

Another line of work in active learning is the study of active learning with surrogate losses. Since optimization under zero-one loss is not possible efficiently, one could approach this issue by using a surrogate loss to ease optimization's computational complexity. An extensive theoretical study of this subject is done by Hanneke (2014) and **?**. There are many other interesting problems studied in active learning that we have not been covered here.

Chapter 3

IWAL-σ

3.1 Introduction

In this chapter we study online active learning under the Bernstein condition for general, bounded losses. The Bernstein condition (see Definition 3.2) is a low noise condition. It has been shown that under the Bernstein condition fast rates of learning are achievable for general losses in passive learning (Massart et al., 2006; van Erven et al., 2015). Also, previous works in active learning such as Hanneke (2009); Koltchinskii (2010); Huang et al. (2015) have studied active learning under an adaptation of the Tsybakov noise condition for zero-one loss and achieved up to a logarithmic label complexity in this special case. However, active learning under a low noise condition has not been sufficiently investigated in the case of general losses. We show improvements in label complexity in the case the risk of the best hypothesis in the hypothesis class is small, for general losses and recover logarithmic label complexity for the special case of zero-one loss. We use refined variance-based concentration inequalities (Freedman's inequality) in the design and analysis of our new algorithm, IWAL-σ, and leverage the Bernstein condition to upper bound the variance. IWAL-σ is based on an algorithm called IWAL proposed by Beygelzimer et al. (2009), who studied online active learning for general, bounded losses under no assumption. However, IWAL is designed for general losses, yet under no Bernstein's condition assumptions the label complexity obtained by IWAL is of the same order as the label complexity obtained by Beygelzimer et al. (2010); Huang et al. (2015)'s algorithms even though these latter algorithms are designed only to be used for zero-one loss.

Contributions. Our work is centered around a new algorithm named IWAL-σ designed to be used under the Bernstein condition and flexible to any bounded loss function. We show how this condition can be used to be beneficial in some cases. We prove both generalization bounds and label complexity results for IWAL-σ. Finally, we study the special case of zero-one loss and achieve up to exponential improvements in label complexity over passive learning under the Tsybakov noise condition.

The rest of the paper is structured as follows. We begin by covering some preliminaries. Before diving into our contribution, we review the Importance Weighted Active Learning algorithm (IWAL) since our work is based on IWAL. Next, we introduce and analyze our algorithm, named IWAL-σ. In Section 3.6, we discuss an adapted version of our algorithm for zero-one loss.

3.2 Preliminaries

In this section, we cover some of the frequently used notation in this paper. We draw samples i.i.d. from an unknown and fixed distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ are input and output spaces respectively. The class of hypotheses is denoted by $\mathcal{H}$, where, for each hypothesis f in $\mathcal{H}$, f maps the input space to the prediction space, $\mathcal{Z} \subseteq \mathbb{R}$. Let a loss function $\ell(z, y)$ be a mapping from $\mathcal{Z} \times \mathcal{Y}$ to $[0, 1]$, where $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$. Denote by $f^* = \arg\min_{f \in \mathcal{H}} R(f)$ the best hypothesis in $\mathcal{H}$, where $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$. In our case of study, online active learning, at the beginning of each round t, Nature provides the learner with a new sample x_t drawn from $\mathcal{D}_x$, where $\mathcal{D}_x$ is the marginal distribution over the input space. Next, the learner decides whether or not to query y_t.

3.3 Review of IWAL

We review IWAL (see Algorithm 1), which is an important algorithm among the active learning community and, most importantly in this work, our algorithm IWAL$-\sigma$ is based on IWAL. The main idea of IWAL is to maintain a good hypothesis set called the effective version space, denoted by $\mathcal{H}_t$, for each round t. At every round t, IWAL maintains an effective version space of hypotheses $\mathcal{H}_{t+1}$ for the next round, where initially $\mathcal{H}_1 = \mathcal{H}$. At the beginning of every round, the algorithm observes a sample x_t. Next, with some probability $p_t(x_t)$, the algorithm will query the label y_t, where

$p_t(x_t)$ is computed as

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

Here, $\mathcal{L}(f(x), g(x))$ is the maximum possible disagreement value between two hypotheses, f and g on point $x \in \mathcal{X}$, defined as

$$\mathcal{L}(f(x), g(x)) = \max_{y \in \mathcal{Y}} |\ell(f(x), y) - \ell(g(x), y)|,$$

and ℓ has range in $[0, 1]$. After observing $p_t(x_t)$, the algorithm draws a sample, Q_i, from a Bernoulli distribution with parameter $p_t(x_t)$. The algorithm queries y_t if and only if Q_i is 1. Whether the algorithm queries y_t or not, it updates the importance weighted loss of each hypothesis $L_t(f)$, defined as

$$L_t(f) = \frac{1}{t} \sum_{i=1}^{t} \frac{Q_i}{p_i(x_i)} \ell(f(x_i), y_i).$$

Finally, any $f \in \mathcal{H}_t$ that satisfies $L_t(f) \leq L_t(\hat{f}_t) + \Delta_t$ will be kept in $\mathcal{H}_{t+1}$, where $\Delta_t = \sqrt{(8/t)\ln(2t(t+1))|\mathcal{H}|^2/\delta}$ and $\hat{f}_t = \arg\min_{f \in \mathcal{H}_t} L_t(f)$. The thresholding-based elimination can be written as

$$\mathcal{H}_{t+1} = \{f \in \mathcal{H}_t : L_t(f) \leq L_t(\hat{f}_t) + \Delta_t\}.$$

Beygelzimer et al. (2009) prove that any hypothesis f kept in $\mathcal{H}_t$ has an excess risk not worse than Δ_t. We refer to Δ_t as upper deviation term.

The original analysis of IWAL (Beygelzimer et al., 2009) was done only for losses with bounded slope asymmetry (See Definition 4 in Beygelzimer et al. (2009)); later, an improved analysis was proposed by Cortes et al. (2019b), which also relaxed the algorithm to use bounded general losses. In Beygelzimer et al. (2009), the label complexity bound grows as $K_l \cdot \theta_{IWAL}$, where K_l is the slope asymmetry and θ_{IWAL} is the disagreement coefficient (See Definition 9 in Beygelzimer et al. (2009)). In Cortes et al. (2019b)'s analysis, they used a different definition for disagreement coefficient θ, (which is the same definition that we use, see Section 3.5 for the definition). They improved the label complexity bound by replacing $K_l \cdot \theta_{IWAL}$ by θ. Later, Zhang (2019) proved that $\theta \leq K_l \cdot \theta_{IWAL}$.

Cortes et al. (2019b) also proposed IWAL-D, which is an improved version of IWAL. They improve IWAL by reducing Δ_t to $\frac{1+\mathcal{L}(f,\hat{f}_t)}{2}\Delta_t$, where $\mathcal{L}(f, g) = \mathbb{E}[\mathcal{L}(f(x), g(x))]$.

The assumption that $\mathcal{L}(f, g)$ is accessible is reasonable since in active learning we can assume that learner has access to a large unlabeled dataset for free. Thus, IWAL-D can compute an estimate of $\mathcal{L}(f, g)$ prior to requesting any label. Even though IWAL-D's improvements might not be large, there is the question of whether it is possible to define a better upper deviation bound to reduce the number of labels queried.

Algorithm 1: IWAL($\mathcal{H}, \delta, T$)

```
$\mathcal{H}_1 = \mathcal{H}$
for $t \in [T]$ do
    Receive $x_t$
    $P_t \leftarrow \max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x_t), g(x_t))$
    $Q_t \leftarrow Bernoulli(P_t)$
    if $Q_t$ then
        $y_t \leftarrow Label(x_t)$
    end
    $\hat{f}_t \leftarrow \arg\min_{f \in \mathcal{H}_t} L_t(f)$
    $\Delta_t = \sqrt{(8/t)\ln(2t(t+1))|\mathcal{H}|^2/\delta}$
    $\mathcal{H}_{t+1} = \{\forall f \in \mathcal{H}_t : L_t(f) \leq L_t(\hat{f}_t) + \Delta_t\}$
end
```

3.4 IWAL-σ

Inspired by previous works in passive learning and the agnostic active learning line of work in active learning, we study the case of agnostic learning under a low noise constraint. To study such problems, first, we must formally define the low noise condition. One of the most commonly used notions of restricted noise in active learning is an adaptation of the Tsybakov noise condition (Mammen and Tsybakov, 1999; Beygelzimer et al., 2010; Huang et al., 2015; Hanneke, 2014).

Definition 3.1 (Tsybakov noise condition)**.** *A learning problem D with hypothesis class $\mathcal{H}$ satisfies the Tsybakov noise condition with exponent $\alpha \in [0, 1]$ and non zero constant C if*

$$P(f(X) \neq f^*(X)) \leq C(R(f) - R(f^*))^{\alpha} \qquad \textit{for all } f \in \mathcal{H}.$$

It is worth mentioning that in the original Tsybakov noise condition (Mammen and Tsybakov, 1999), $R(f^*_{Bayes})$ is used instead of $R(f^*)$, where f^*_{Bayes} or Bayes optimal learner is the best possible predictor function. Assuming $f^*_{Bayes} \in \mathcal{H}$ is a strong assumption and hence not desired. This likely is one of the main reasons previous works such as Huang et al. (2015); Beygelzimer et al. (2010) considered to use an adapted version of the original definition of the Tsybakov noise condition. Here we use the Bernstein condition; these two notions look quite similar and yet are fundamentally different. One of the differences is that the Bernstein condition applies to general losses; however, the commonly used Tsybakov noise condition is only applicable in the case of zero-one loss.

Definition 3.2 (Bernstein condition)**.** *A learning problem D with hypothesis class $\mathcal{H}$ satisfies the $(C, \beta)-$Bernstein condition if*

$$\mathbb{E}_{x,y\sim D}[(\ell(f(x), y) - \ell(f^*(x), y))^2] \leq C(R(f) - R(f^*))^\beta \qquad \textit{for all } f \in \mathcal{H},$$

where C is a non zero constant and $0 \leq \beta \leq 1$.

Our goal is to introduce an algorithm that benefits from the Bernstein condition. To do so, we use the variance of importance weighted losses in our generalization bound, which helps us to achieve a tighter bound. Before discussing the algorithm and the details, let us review and redefine a few definitions. The probability of querying some x at round t is denoted by

$$p_t(x) = \max_{f,g\in\mathcal{H}_t} \mathcal{L}(f(x), g(x)); \tag{3.1}$$

in particular, let $P_t := p_t(x_t)$, where x_t is the sample observed on round t. Let $Q_t(x)$ be a sample from $Bernoulli(p_t(x))$ and Q_t be sample drawn from $Bernoulli(P_t)$. Like before, the importance weighted loss is defined as $L_t(f) = \frac{1}{t}\sum_{i=1}^{t} \frac{Q_i}{P_i}\ell(f(x_i), y_i)$. Most importantly

$$Z_{f,f^*,t} = \begin{cases} \frac{Q_t}{P_t}\left(\ell(f(x_t), y_t) - \ell(f^*(x_t), y_i)\right) - (R(f) - R(f^*)) & \text{if } f, g \in \mathcal{H}_t \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{H}_t$ is the effective version space at round t. Conceptually, $\mathcal{H}_t$ is similar to the effective version space explained before; however, the details of maintaining $\mathcal{H}_t$ in IWAL-σ are different and will be covered after introducing a few more definitions.

The upper deviation term $\Delta_{f,g,t}$ is defined as

$$\Delta_{f,g,t} = \frac{1}{t}\max\left(\sqrt{4\sigma^2_{f,g,t}\ln(\frac{4\ln(t)}{\delta'_t}}), 6\ln({}^{4\ln(t)}/{}_{\delta'_t})\right),$$

where $f,g \in \mathcal{H}$, t is the round number, $\sigma^2_{f,g,t} = \sum_{i=1}^{t} Var(Z_{f,g,i} \mid \mathcal{F}_{i-1})$, and δ'_t is the confidence variable. Also, the history is denoted by $\mathcal{F}_t = \{A_1, A_2, \ldots, A_t\}$, where $A_i = \{x_i, y_i, Q_i\}$. The effective version space at the beginning of round t is denoted by

$$\mathcal{H}_t = \{f \in \mathcal{H}_{t-1} : L_t(f) \leq L_t(\hat{f}_t) + \Delta_{f,\hat{f}_t,t}\},$$

where $\hat{f}_t = \arg\min_{f\in\mathcal{H}_t} L_t(f)$.

In the rest of this section, we first (in Lemmas 3.1 and 3.2) prove a generalization bound. Next we discuss why we cannot use $\Delta_{f,g,t}$ directly in IWAL-σ, suggest an alternative, $\hat{\Delta}_{f,g,t}$, and then rewrite the results of Lemmas 3.1 and 3.2 based on $\hat{\Delta}_{f,g,t}$. Finally, we study label complexity.

Lemma 3.1. *For all probability distributions D, for all hypothesis classes $\mathcal{H}$, for all $\delta > 0$, with probability at least $1-\delta$, for all T and any $f, g \in \mathcal{H}_T$,*

$$|L_T(f) - L_T(g) - R(f) + R(g)| \leq \Delta_{f,g,T},$$

where $\Delta_{f,g,T} = \frac{1}{T}\max\left(\sqrt{4\sigma^2_{f,g,T}\ln(\frac{4\ln(T)}{\delta'_T}}), 6\ln\left(\frac{4\ln(T)}{\delta'_T}\right)\right)$, $\delta'_T = \frac{\delta}{|\mathcal{H}|^2T(T+1)}$, *and* $\sigma^2_{f,g,T} = \sum_{t=1}^{T} Var(Z_{f,g,t} \mid \mathcal{F}_{t-1})$.

Proof. In this proof, we fix $f, g \in \mathcal{H}$ and for the sake of readability, we summarize $Z_{f,g,t}$ by Z_t. Then by the law of total expectation, we can write

$$\begin{aligned}
\mathbb{E}\left[Z_t \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbf{1}\left[f,g\in\mathcal{H}_t\right]\left(\frac{Q_t}{P_t}(\ell(f(x_t),y_t) - \ell(g(x_t),y_t)) - R(f) + R(g)\right) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[f,g\in\mathcal{H}_t\right](\ell(f(x_t),y_t) - \ell(g(x_t),y_t) - R(f) + R(g)) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbf{1}\left[f,g\in\mathcal{H}_t\right]\mathbb{E}\left[(\ell(f(x_t),y_t) - \ell(g(x_t),y_t)) \mid \mathcal{F}_{t-1}\right] - R(f) + R(g) = 0
\end{aligned}$$

Therefore, $Z_1, Z_2, \ldots$ is a martingale difference sequence for any fixed f, g. We can use Freedman's inequality to bound the sum of Z_t, hence $Z_t \leq 2$ where the loss function is ranged in $[0, 1]$. Also note that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} Z_t &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[f, g \in \mathcal{H}_t\right]\left(\frac{Q_t}{P_t}(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)) - R(f) + R(g)\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{Q_t}{P_t}(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)) - R(f) + R(g)\right) \\ &= L_T(f) - L_T(g) - (R(f) - R(g))\end{aligned}$$

where the second equality holds if $f, g \in \mathcal{H}_T$. We like to bound $\Pr[|\sum_{t=1}^{T} Z_t| \geq T\Delta_{f,g,T}]$ for a fixed T, f and g. By applying Freedman's inequality (see Lemma A.1) from Kakade and Tewari (2009), we have

$$\Pr\left[\sum_{t=1}^{T} Z_t \geq \max\left(2\sigma_{f,g,t}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}, 6\ln(\frac{4\ln(T)}{\delta'_T})\right)\right] \leq \delta'_T.$$

Taking a union bound over all $f, g \in \mathcal{H}$ and another union bound over T completes the proof. □

Lemma 3.2. *For any probability distribution D and hypothesis class $\mathcal{H}$, let $f^* \in \mathcal{H}$ be a minimizer of the loss function with respect to D. For any $\delta > 0$, with probability at least $1 - \delta$: (i) $f^* \in \mathcal{H}_t$ for any t, and (ii) $R(\hat{f}_t) - R(f^*) \leq \Delta_{f^*,\hat{f}_t,t}$ for any $t \geq 2$.*

Proof. Similar to Beygelzimer et al. (2009) we use mathematical induction on t to prove the first part. The base case trivially holds for $t = 1, 2$ hence $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$ and $f^* \in \mathcal{H}$, this holds since shrinking the effective version space can be started at second round. We assume that this claim holds for $t = T$ for some $T > 2$. We are going to prove it for $t = T + 1$. Using Lemma 3.1, where $f = f^*$, we can write

$$L_t(f^*) - L_t(\hat{f}_t) \leq \Delta_{f^*,\hat{f}_t,t} + R(f^*) - R(\hat{f}_t) \leq \Delta_{f^*,\hat{f}_t,t}, \tag{3.2}$$

where the first inequality holds because of Lemma 3.1 with the second inequality holds because $R(f^*) - R(\hat{f}_t)$ is non-positive. By moving $L_t(\hat{f}_t)$ we can write

$$L_t(f^*) \leq L_t(\hat{f}_t) + \Delta_{f^*,\hat{f}_t,t}. \tag{3.3}$$

This proves the algorithm keeps f^* in $\mathcal{H}_{t+1}$. To prove the second part, we use Lemma

3.1 again.

$$R(\hat{f}_t) - R(f^*) \leq \Delta_{f^*,\hat{f}_t,t} - L_t(f^*) + L_t(\hat{f}_t) \leq \Delta_{f^*,\hat{f}_t,t}$$

where the second inequality holds because $L_t(f^*) \geq L_t(\hat{f}_t)$. □

While Lemma 3.2 provides a generalization bound, the bound is not descriptive enough since it depends on $\sigma^2_{f^*,\hat{f}_t,t}$, which might be small or large. Thus, to easily interpret the generalization bound, we need to upper bound $\sigma^2_{f^*,\hat{f}_t,t}$, which will be studied in Lemma 3.3. Besides not being easily interpretable, we do not know $\sigma^2_{f,g,t}$ in advance, and so it is not possible to directly use $\Delta_{f,g,t}$ in an algorithm. Therefore, we need to estimate the variance of $Z_{f,g,t}$. To do so, we use Lemma 3.4. We start by bounding $\sigma^2_{f,f^*,T}$, after which we approach the second problem.

Lemma 3.3. *Under the assumption that the (C, β)-Bernstein condition holds, for any $f \in \mathcal{H}$, $\sigma^2_{f,f^*,T} \leq T\sqrt{C(R(f) - R(f^*))^\beta}$.*

Proof. Recall that,

$$\sigma^2_{f,f^*,T} = \sum_{t=1}^{T} Var(Z_{f,f^*,t} \mid \mathcal{F}_{t-1})$$

$$Z_{f,f^*,t} = \begin{cases} \frac{Q_t}{P_t}\left(\ell(f(x_t), y_t) - \ell(f^*(x_t), y_i)\right) - (R(f) - R(f^*)) & \text{if } f, g \in \mathcal{H}_t \\ 0 & \text{otherwise} \end{cases}$$

By definition, we have

$$
\begin{aligned}
&\sigma^2_{f,f^*,T} = \sum_{t=1}^{T} \left(\mathbb{E}\left[Z^2_{f,f^*,t} \mid \mathcal{F}_{t-1} \right] - \mathbb{E}\left[Z^2_{f,f^*,t} \mid \mathcal{F}_{t-1} \right]^2 \right) \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[Z^2_{f,f^*,t} \mid \mathcal{F}_{t-1} \right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \left(\frac{Q_t}{P_t} (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t)) - (R(f) - R(f^*)) \right)^2 \mid \mathcal{F}_{t-1} \right] \\
&= \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \mathbb{E}\left[\left(\frac{Q_t}{P_t} (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t)) - (R(f) - R(f^*)) \right)^2 \mid \mathcal{F}_{t-1} \right] \\
&\overset{(\spadesuit)}{=} \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \left(\mathbb{E}\left[\left(\frac{Q_t}{P_t} (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t)) \right)^2 \mid \mathcal{F}_{t-1} \right] - (R(f) - R(f^*))^2 \right) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \mathbb{E}\left[\left(\frac{Q_t}{P_t} (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t)) \right)^2 \mid \mathcal{F}_{t-1} \right] \\
&\overset{(\heartsuit)}{\leq} \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \mathbb{E}\left[\frac{Q_t}{P_t} \left| \ell(f(x_t), y_t) - \ell(f^*(x_t), y_t) \right| \mid \mathcal{F}_{t-1} \right] \\
&\overset{(\clubsuit)}{=} \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \mathbb{E}\left[\left| \ell(f(x_t), y_t) - \ell(f^*(x_t), y_t) \right| \mid \mathcal{F}_{t-1} \right] \\
&= \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \mathbb{E}\left[\sqrt{(\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t))^2} \mid \mathcal{F}_{t-1} \right] \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \sqrt{\mathbb{E}\left[(\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t))^2 \mid \mathcal{F}_{t-1} \right]} \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left[f, f^* \in \mathcal{H}_t\right] \sqrt{C(R(f) - R(f^*))^\beta} = T\sqrt{C(R(f) - R(f^*))^\beta}.
\end{aligned}
$$

We briefly explain why the non-trivial steps above hold. First, we argue why each of equalities/inequalities marked by $(\spadesuit)$, $(\heartsuit)$, and $(\clubsuit)$ holds; these all hold "point-wise" (for each $t \in [T]$), and so we consider a fixed t. In the case that $\mathbf{1}\left[f, f^* \in \mathcal{H}_t\right]$ is zero, they all hold as then the LHS and RHS for each is equal to zero. Therefore, we now consider the case that $f, f^* \in \mathcal{H}_t$. In this case, $\overset{(\spadesuit)}{=}$ holds because $\mathbb{E}\left[\frac{Q_t}{P_t} (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t)) \mid \mathcal{F}_{t-1} \right] = R(f) - R(f^*)$; next, $\overset{(\heartsuit)}{\leq}$ holds because $P_t \geq (\ell(f(x_t), y_t) - \ell(f^*(x_t), y_t))$; last, $\overset{(\clubsuit)}{=}$ holds by taking the conditional expectation of Q_t. The line after that follows by applying Jensen's inequality, and the final line uses the Bernstein condition. □

By substituting $\sigma^2_{\hat{f}_T,f^*,T}$ by $T\sqrt{C(R(\hat{f}_T) - R(f^*))^\beta}$ into $\Delta_{f^*,\hat{f}_T,T}$ and using part (ii) of Lemma 3.2, we can finally see that

$$R(\hat{f}_T) - R(f^*) \leq O\left(\frac{1}{T}\sqrt{C}\ln(4\ln(T)/\delta'_T)\right)^{2/(4-\beta)}.$$

Next, we attempt to estimate $\sigma^2_{f,g,t}$ by Lemma 3.4 below in order to obtain a new, empirical version of $\Delta_{f,g,T}$, later denoted by $\hat{\Delta}_{f,g,T}$. This lemma estimates $\sigma^2_{f,g,t}$ by $\hat{V}_{f,g,t}$, where $\hat{V}_{f,g,t}$ is an observable estimate of $\sigma^2_{f,g,t}$ (see (3.7) for details). Given $\hat{V}_{f,g,t}$, we estimate $\Delta_{f,g,t}$ by $\hat{\Delta}_{f,g,t}$ and use $\hat{\Delta}_{f,g,t}$ in IWAL-σ. Lemma 3.4 is inspired by Lemma 3 from Peel et al. (2013); however, we had to make changes to their result to utilize it in our algorithm.

Lemma 3.4. *Let $\{X_t\}_{t\in[T/2]}$ be a stochastic process adapted to a filtration $\{\mathcal{G}_t\}_{t\in[T/2]}$, where, for each $t \in \{0, 1, \ldots, T/2-1\}$, X_{2t+1} and X_{2t+2} are conditionally i.i.d. given $X_1, \ldots, X_{2t}$ (i.e., given $\mathcal{G}_{2t}$). We use the notation $g_{\mathcal{G}_t}$ in order to indicate a random function that, ignoring its argument, is $\mathcal{G}_t$-measurable; therefore, such a function is fixed after time t. Suppose $\{g_{\{\mathcal{G}_t\}}\}_{t\in[T]}$ is a family of functions which take their values in $[0, 1/2]$; then for all $0 < \delta_\sigma \leq 1$*

$$\Pr\left[|\hat{V}_T - V_T| \geq \sqrt{\frac{T}{4}\ln\left(\frac{2}{\delta_\sigma}\right)}\right] \leq \delta_\sigma$$

where

$$V_T = \sum_{t=1}^{T/2} Var\left[g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) + g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}) \mid \mathcal{G}_{2t-2}\right]$$

and

$$\hat{V}_T = \sum_{t=1}^{T/2}(g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) - g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}))^2. \tag{3.4}$$

Proof. First, we observe that

$$Var\left[g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) + g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}) \mid \mathcal{G}_{2t-2}\right] = 2Var\left[g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}) \mid \mathcal{G}_{2t-2}\right]. \tag{3.5}$$

Next, we turn to forming an estimator for these conditional variances. Define M_t as

$$M_t := (g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) + g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}))^2.$$

Note that M_t is $\mathcal{G}_{2t}$-measurable. Next, define B_t as

$$B_t := M_t - \mathbb{E}\left[M_t \mid \mathcal{G}_{2t-2}\right].$$

Next, we observe that the sequence $\{B_t\}_{t\geq 1}$ is a martingale difference sequence. To see this clearly, we define $\mathcal{G}'_t := \mathcal{G}_{2t}$ for each $t \in [T/2]$. Then we can re-express M_t and B_t as $M_t = (g_{\{\mathcal{G}'_{t-1}\}}(X_{2t-1}) + g_{\{\mathcal{G}'_{t-1}\}}(X_{2t}))^2$ and $B_t = M_t - \mathbb{E}\left[M_t \mid \mathcal{G}'_{t-1}\right]$ respectively, each of which is $\mathcal{G}'_t$-measurable. Finally, since $\mathbb{E}\left[B_t \mid \mathcal{G}'_{t-1}\right] = 0$, the sequence $\{B_t\}_{t\geq 1}$ is indeed a martingale difference sequence.

Since each function in the family is $[0, 1/2]$-valued, we have $B_t \in [-1/2, 1/2]$. Also, observe that the sequence $\left\{\sum_{t=1}^{T} B_t\right\}_{T\geq 1}$ is a martingale. Applying the Azuma-Hoeffding inequality on this sequence yields

$$\Pr\left(\sum_{t=1}^{T/2} B_t \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{T/2}\right)$$

Finally observe that

$$\begin{aligned}\mathbb{E}\left[M_t|\mathcal{G}_{2t-2}\right] &= \mathbb{E}\left[(g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) + g_{\{\mathcal{G}_{2t-2}\}}(X_{2t}))^2 \mid \mathcal{G}_{2t-2}\right] \\ &= 2Var\left[g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) \mid \mathcal{G}_{2t-2}\right],\end{aligned}$$

matching (3.5). Hence we have

$$Pr\left(|\hat{V}_T - \sum_{t=1}^{T/2} 2Var\left[g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) \mid \mathcal{G}_{2t-2}\right]| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{T/2}\right).$$

The proof is concluded by setting $\epsilon = \sqrt{\frac{T}{4}\ln\left(\frac{2}{\delta^\sigma_T}\right)}$. □

To use Lemma 3.4 we need to clarify what are X_t and $g_{\{\mathcal{G}_t\}}$ in our case. Let A_t be

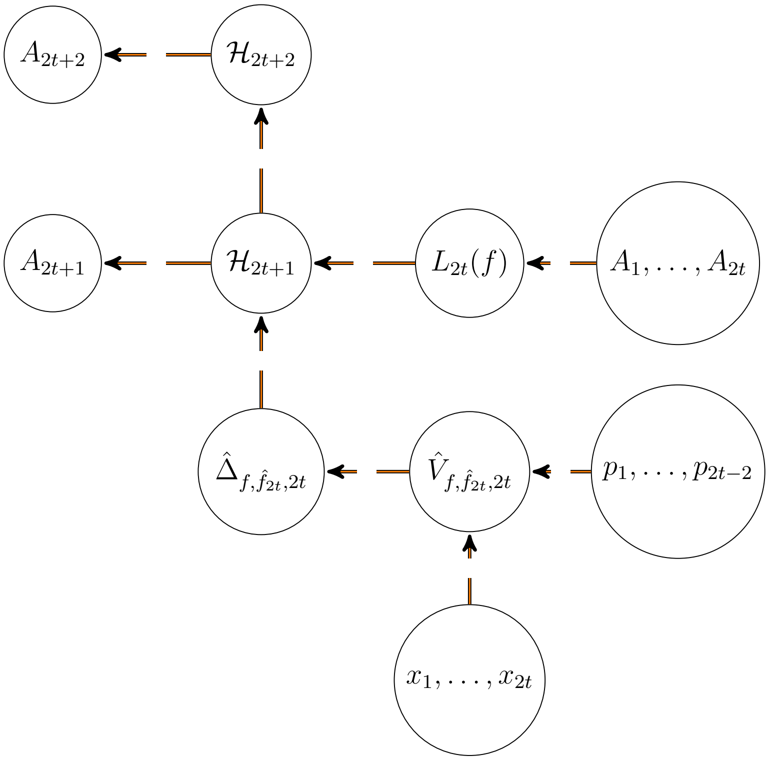


Figure 3.1: Dependency Graph

X_t (in the lemma). We set

$$\begin{aligned}
g_{\{\mathcal{G}_{2t-2}\}}(X_{2t-1}) &= g_{\{\mathcal{F}_{2t-2}\}}(A_{2t-1}) \\
&= \mathbf{1}\left[f, g \in \mathcal{H}_{2t-1}\right] \Bigg(\frac{Q_{2t-1}(x_{2t-1})}{4p_{2t-1}(x_{2t-1})} (\ell(f(x_{2t-1}), y_{2t-1}) - \ell(g(x_{2t-1}), y_{2t-1})) \\
&\qquad - \frac{1}{4}(R(f) - R(g)) \Bigg)
\end{aligned}$$

where f and g are fixed functions from $\mathcal{H}$. We have to make sure that A_{2t+1} and A_{2t+2} only depend on $A_1, A_2, \ldots, A_{2t}$. Figure 3.1 demonstrates dependencies of A_{2t+1} and A_{2t+2}. As shown in the figure, A_{2t+1} and A_{2t+2} depend on $A_1, A_2, \ldots, A_{2t}$. So, in our case we can write V_T as

$$V_{f,g,T} = 16 \sum_{t=1}^{T/2} Var\left[Z_{f,g,2t-1}/4 + Z_{f,g,2t}/4 | \mathcal{F}_{2t-2}\right] = 16 \sum_{t=1}^{T/2} 2Var\left[Z_{f,g,2t-1}/4 | \mathcal{F}_{2t-2}\right] \tag{3.6}$$

and, for even T, we can write $\hat{V}_T$ as

$$\hat{V}_{f,g,T} = 16 \sum_{t=1}^{T/2} \mathbf{1}\left[f, g \in \mathcal{H}_{2t-1}\right] \left(\frac{Q_{2t-1}(x_{2t-1})}{4p_{2t-1}(x_{2t-1})} \ell_{f-g}(x_{2t-1}, y_{2t-1}) - \frac{Q_{2t}(x_{2t})}{4p_{2t}(x_{2t})} \ell_{f-g}(x_{2t}, y_{2t}) \right)^2, \tag{3.7}$$

where we use the fact that $\mathcal{H}_{2t-1} = \mathcal{H}_{2t}$ and use the notation $\ell_{f-g}(x, y) = \ell(f(x), y) - \ell(g(x), y)$. Denote by $\hat{\Delta}_{f,\hat{f}_T,T}$ an adapted version of $\Delta_{f,\hat{f}_T,T}$ by replacing $\sigma^2_{f,g,T}$ with $\hat{V}_{f,g,T} + 16\sqrt{(T/4)\ln(2/\delta)}$. In the next result, since the claim is only for $f, g \in \mathcal{H}_T$, the indicator $\mathbf{1}\left[f, g \in \mathcal{H}_{2t-1}\right]$ in (3.7) is always one and hence can be ignored.

Corollary 3.1. *For all probability distributions D, for all hypothesis classes $\mathcal{H}$, for all $\delta > 0$, with probability at least $1 - \delta$, for all T and any $f, g \in \mathcal{H}_T$,*

$$|L_T(f) - L_T(g) - R(f) + R(g)| \leq \hat{\Delta}_{f,g,T}$$

where

$$\hat{\Delta}_{f,g,T} = \frac{1}{T} \max\left(2\sqrt{\hat{V}_{f,g,T} + 16\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)}} \sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}, 6\ln(\frac{4\ln(T)}{\delta'_T}) \right)$$

and $\delta'_T = \delta^\sigma_T = \frac{\delta}{2|\mathcal{H}|^2 T(T+1)}$.

Proof. The proof is the same as Lemma 3.1 except that we replace $\sigma^2_{f,g,T}$ by $\hat{V}_{f,g,T} + 16\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)}$. Since $\hat{V}_{f,g,T} + 16\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)}$ is larger than $\sigma^2_{f,g,T}$, the proof is still valid. □

Theorem 3.1. *For any probability distribution D and hypothesis class $\mathcal{H}$, let $f^* \in \mathcal{H}$ be a minimizer of the loss function with respect to D. For any $\delta > 0$, with probability at least $1 - \delta$: (i) $f^* \in \mathcal{H}_t$ for any t, and (ii) $R(\hat{f}_t) - R(f^*) \leq \hat{\Delta}_{f^*,\hat{f}_t,t}$ for any $t \geq 2$ (iii) for any $f \in \mathcal{H}_t$ and $t \geq 2$ we have $R(f) - R(f^*) \leq \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}$, where $\tilde{\Delta}_t = \max_{f \in \mathcal{H}_t} \hat{\Delta}_{f,\hat{f}_t,t}$.*

Proof. The proof for (i) and (ii) are the same as the proof of Lemma 3.2 but using Corollary 3.1's results instead of Lemma 3.1's results. To prove the third part, first observe that from Corollary 3.1 for any $f \in \mathcal{H}_t$ we can write

$$\begin{aligned} L_{t-1}(\hat{f}_t) - L_{t-1}(f) - R(\hat{f}_t) + R(f) &\leq \hat{\Delta}_{f,\hat{f}_t,t-1} \Leftrightarrow \\ R(f) - R(\hat{f}_t) &\leq \hat{\Delta}_{f,\hat{f}_t,t-1} - L_{t-1}(\hat{f}_t) + L_{t-1}(f). \end{aligned} \tag{3.8}$$

Next, we bound $\hat{\Delta}_{f,\hat{f}_t,t-1}$ by $\tilde{\Delta}_{t-1}$ and $L_{t-1}(f) - L_{t-1}(\hat{f}_t)$ by $\tilde{\Delta}_{t-1}$ since $f, \hat{f}_t \in \mathcal{H}_t$ and we get

$$R(f) - R(\hat{f}_t) \leq \hat{\Delta}_{f,\hat{f}_t,t-1} + \hat{\Delta}_{f,\hat{f}_t,t-1} = 2\tilde{\Delta}_{t-1}. \tag{3.9}$$

By putting together (3.9) and the fact that $R(\hat{f}_t) - R(f^*) \leq \tilde{\Delta}_t$, we have

$$R(f) - R(f^*) = (R(\hat{f}_t) - R(f^*)) + (R(f) - R(\hat{f}_t)) \leq \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}.$$

□

Corollary 3.1 and Theorem 3.1 together prove a generalization bound with respect to $\hat{\Delta}_{f^*,\hat{f}_T,T}$. The next corollary demonstrates the generalization bound with respect to only δ and T.

Corollary 3.2. *Under the assumption that the Bernstein condition holds,* $R(\hat{f}_T) - R(f^*) \leq O\left(\frac{1}{T}\sqrt{C}\ln(4\ln(T)/\delta'_T)\right)^{2/(4-\beta)}$.

Proof. First, we know that

$$R(\hat{f}_T) - R(f^*) \leq \hat{\Delta}_{f^*,\hat{f}_T,T}. \tag{3.10}$$

Next, we upper bound $\hat{V}_{f^*,\hat{f}_T,T}$ which is used in $\hat{\Delta}_{f^*,\hat{f}_T,T}$ with

$$\begin{aligned}\hat{V}_{f^*,\hat{f}_T,T} &\leq \sigma^2_{f,g,T} + 16\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)} \\ &\leq T\sqrt{C(R(\hat{f}_T) - R(f^*))^\beta} + 16\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)},\end{aligned} \tag{3.11}$$

where the first inequality is a direct application of Lemma 3.4 and the second inequality is derived by upper bounding $\sigma^2_{f,g,T} \leq T\sqrt{C(R(\hat{f}_T) - R(f^*))^\beta}$. By upper bounding $\hat{V}_{f^*,\hat{f}_T,T}$ in $\hat{\Delta}_{f^*,\hat{f}_T,T}$ in (3.10), we get the following:

$$\begin{aligned}&R(\hat{f}_T) - R(f^*) \\ &\leq \frac{1}{T}\max\left(2\sqrt{T\sqrt{C(R(\hat{f}_T) - R(f^*))^\beta} + 32\sqrt{\frac{T}{4}\ln(2/\delta^\sigma_T)}}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}, 6\ln\left(\frac{4\ln(T)}{\delta'_T}\right)\right).\end{aligned}$$

To solve this equation, first we use

$$\sqrt{T\sqrt{C(R(\hat{f}_T)-R(f^*))^\beta}+32\sqrt{\frac{T}{4}}\ln(2/\delta_T^\sigma)} \leq 2\max(\sqrt{T\sqrt{C(R(\hat{f}_T)-R(f^*))^\beta}}, \sqrt{32\sqrt{\frac{T}{4}}\ln(2/\delta_T^\sigma)})$$

and then solve the equation. Since the results based on the term $\sqrt{T\sqrt{C(R(\hat{f}_T)-R(f^*))^\beta}}$ is larger for any $0 \leq \beta \leq 1$ we avoid the max in the final result. □

3.5 Label Complexity

As we discussed earlier, to propose an active learning algorithm, we should provide a generalization bound and an upper bound on the number of labels queried by the algorithm. In this section, we discuss the latter. Before upper bounding the number of queried samples, we have to introduce the notion of disagreement coefficient. The disagreement coefficient, denoted by θ, is the infimal value that satisfies the following for any r:

$$\mathbb{E}_{x\sim D}[\max_{f\in B(f^*,r)} \mathcal{L}(f(x), f^*(x))] \leq \theta r,$$

where

$$B(f,r) := \{g \in \mathcal{H} : D(f,g) \leq r\}$$

and

$$D(f,g) := \mathbb{E}\left[|\ell(f(x),y) - \ell(g(x),y)|\right].$$

The disagreement coefficient is a commonly used notion among the active learning community. It was first introduced by Hanneke (2007). Later Beygelzimer et al. (2009) has extended this definition for general losses. This particular definition used in this work was recently introduced by Cortes et al. (2019a). Hanneke (2007, 2014) has bounded the value of θ is different cases such as linear separators under uniform distribution, and more generally, the value of θ for zero-one loss. Throughout this section δ_T^σ, δ'_T, and δ are used as confidence variables and are defined as $\delta_\sigma^T = \delta'_T = \frac{\delta}{2|\mathcal{H}|^2T(T+1)}$, where δ is the confidence variable used in Corollary 3.1. We show two different upper bounds for label complexity. First, in Corollary 3.4 we provide a fallback guarantee by showing that the label complexity is not worse than θ/ϵ, where ϵ

is the generalization bound. This shows that our algorithm is not more expensive than passive learning techniques in terms of label complexity. Unfortunately, we suspect it is not possible to improve this result further in general (Hanneke and Yang, 2010), and thus we focus on two special cases. First, we focus on situations where $R(f^*)$ is small and prove in such cases it is possible to improve the label complexity down to $\theta/\sqrt{\epsilon}$. Second, we study the case of zero-one loss in Section 3.6, in which case the Bernstein condition becomes the Tsybakov noise condition, and decrease the label complexity down to $\log(\theta/\epsilon)$, a result that was previously achieved by Hanneke (2009); Koltchinskii (2010); Huang et al. (2015).

Theorem 3.2. *Under the assumption that the Bernstein condition holds, for any $\delta > 0$, with probability $1-\delta$, for any $t \geq 2$, the following holds for the label requesting indicator P_t of IWAL-σ:*

$$\mathbb{E}_{x\sim\mathcal{D}_x}\left[P_t \mid \mathcal{F}_{t-1}\right] \leq 2\theta r_t,$$

where $r_t = \sqrt{C(\tilde{\Delta}_t + 2\tilde{\Delta}_{t-1})^\beta}$ and $\tilde{\Delta}_t = \max_{f\in\mathcal{H}_t}(\hat{\Delta}_{f,f^*,t})$.

Proof. First we start by bounding $D(f,f^*)$ for any $f \in \mathcal{H}_t$:

$$\begin{aligned}\mathbb{E}\left[|\ell(f(x),y)-\ell(f^*(x),y)|\right] &= \mathbb{E}\left[\sqrt{(\ell(f(x),y)-\ell(f^*(x),y))^2}\right] && (3.12)\\ &\leq \sqrt{\mathbb{E}\left[(\ell(f(x),y)-\ell(f^*(x),y))^2\right]} && (3.13)\\ &\leq \sqrt{C(R(f)-R(f^*))^\beta} \leq \sqrt{C(\tilde{\Delta}_t + 2\tilde{\Delta}_{t-1})^\beta} && (3.14)\end{aligned}$$

where in the first inequality we used Jensen's inequality. To derive the second inequality, the Bernstein condition is used, and the last inequality holds with high probability $1-\delta$ for all $f \in \mathcal{H}_t$ and $t \geq 2$ simultaneously due to Theorem 3.1. Second, we prove that $\mathcal{H}_t \subseteq B(f^*, r_t)$ with probability $1-\delta$ for any $t \geq 2$. For a fixed $f, t \geq 2$ and $\delta'_t = \frac{\delta}{|\mathcal{H}|t(t+1)}$ we can write this as

$$\begin{aligned}P\left(f \in \mathcal{H}_t \wedge f \notin B(f^*, r_t)\right) =& P\left(f \in \mathcal{H}_t \wedge D(f,f^*) \geq r_t\right)\\ =& P\left(f \in \mathcal{H}_t \wedge D(f,f^*) \geq \sqrt{C(\tilde{\Delta}_t + 2\tilde{\Delta}_{t-1})^\beta}\right)\\ \leq& P\left(f \in \mathcal{H}_t \wedge R(f) - R(f^*) \geq \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}\right) \leq \delta'_t\end{aligned}$$

where the first two equalities hold by definition and the first inequality holds due to (3.14) and finally Theorem 3.1 is used. By taking a union bound over all $f \in \mathcal{H}_t$ we have

$$P(\exists t \geq 2, \mathcal{H}_t \not\subseteq B(f^*, r_t)) \leq \delta.$$

Next, we can write

$$\begin{aligned}
\mathbb{E}[P_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}[\max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x), g(x)) \mid \mathcal{F}_{t-1}] && (3.15)\\
&\leq 2\mathbb{E}[\max_{f \in \mathcal{H}_t} \mathcal{L}(f(x), f^*(x)) \mid \mathcal{F}_{t-1}] && (3.16)\\
&\leq 2\mathbb{E}[\max_{f \in B(f^*, r_t)} \mathcal{L}(f(x), f^*(x)) \mid \mathcal{F}_{t-1}] \leq 2\theta r_t. && (3.17)
\end{aligned}$$

The following shows why the first inequality holds.

$$\begin{aligned}
\max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x), g(x)) &= \max_{f,g \in \mathcal{H}_t} \max_{y \in Y} |\ell(f(x), y) - \ell(g(x), y)|\\
&= \max_{f,g \in \mathcal{H}_t} \max_{y \in Y} |\ell(f(x), y) - \ell(f^*(x), y) + \ell(f^*(x), y) - \ell(g(x), y)|\\
&\leq \max_{f \in \mathcal{H}_t} \max_{y \in Y} |\ell(f(x), y) - \ell(f^*(x), y)| + \max_{f \in \mathcal{H}_t} \max_{y \in Y} |\ell(f^*(x), y) - \ell(f(x), y)|\\
&\leq 2 \max_{f \in \mathcal{H}_t} \mathcal{L}(f(x), f^*(x)).
\end{aligned}$$

□

Next, because $\tilde{\Delta}_t$ depends on the variance, we should rewrite it in a form that does not depend on variance or excess risk. We already have all the tools to do that.

Corollary 3.3. *Under the assumption that the Bernstein condition holds,*

$$\tilde{\Delta}_T \leq O\left(\left(\frac{C^{1/4}\sqrt{\ln(\frac{\ln(T)}{\delta'_T})}}{\sqrt{T}}\right)^{1+\frac{\beta}{4-\beta}} + \frac{\ln(2/\delta^{\sigma}_T)^{1/4}}{T^{3/4}}\sqrt{\ln(\frac{\ln(T)}{\delta'_T})}\right),$$

where $\tilde{\Delta}_T = \max_{f \in \mathcal{H}_T}(\hat{\Delta}_{f,f^*,T}))$.

Proof. Observe that

$$\begin{aligned}
\tilde{\Delta}_T = \max_{f\in\mathcal{H}_T} \hat{\Delta}_{f,f^*,T} &= \max_{f\in\mathcal{H}_T} \frac{2\sqrt{\hat{V}_{f,f^*,T} + 16\sqrt{\ln(2/\delta_T^\sigma)(T/4)}}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}}{T} \\
&\le \max_{f\in\mathcal{H}_T} \frac{2\sqrt{\sigma^2_{f,f^*,T} + 32\sqrt{\ln(2/\delta_T^\sigma)(T/4)}}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}}{T} \\
&\le \max_{f\in\mathcal{H}_T} \frac{2\sqrt{T\sqrt{C(R(f)-R(f^*))^\beta} + 32\sqrt{\ln(2/\delta_T^\sigma)(T/4)}}\sqrt{\ln(\frac{4\ln(T)}{\delta'})}}{T} \\
&\le O\left(\frac{\sqrt{T\sqrt{C\left(\frac{\sqrt{C}\ln(4\ln(T)/\delta'_T)}{T}\right)^{2\beta/(4-\beta)}} + \sqrt{T\ln(1/\delta_T^\sigma)}}\sqrt{\ln(\frac{\ln(T)}{\delta'_T})}}{T}\right) \\
&\le O\left(\left(\frac{C^{1/2}\ln(\frac{\ln(T)}{\delta'_T})}{T}\right)^{\frac{2}{4-\beta}} + \frac{\ln(2/\delta_T^\sigma)^{1/4}}{T^{3/4}}\sqrt{\ln(\frac{\ln(T)}{\delta'_T})}\right),
\end{aligned}$$

where in the first inequality, we upper bounded $\hat{V}_{f,f^*,T}$ by $\sigma^2_{f,f^*,T} + 16\sqrt{\ln(2/\delta_T^\sigma)(T/4)}$, since the result of Lemma 3.4 is two-sided. The second inequality holds by Lemma 3.3, and the third inequality by the bound on excess risk. □

The next corollary implies a label complexity fallback guarantee.

Corollary 3.4. *The number of labels queried by IWAL-σ after T rounds is*

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[P_t \mid \mathcal{F}_{t-1}] &= \sum_{t=1}^{T} 2\theta C^{1/2}\tilde{O}\left(\left(\frac{1}{\sqrt{t}}\left(C^{1/4}\sqrt{\ln(\frac{\ln(t)}{\delta'_t})}\right)\right)^{\frac{\beta}{4-\beta}}\right) \\
&= 2\theta C^{\frac{1}{2}+\frac{\beta}{8-2\beta}}\tilde{O}\left(T^{1-\frac{\beta}{4-\beta}}\right)
\end{aligned}$$

Using the above result and Corollary 3.2 we can see that the label complexity of IWAL-σ is $\tilde{O}\left(\theta C^{\frac{1}{2}+\frac{\beta}{8-2\beta}}\epsilon^{-(2-\beta)}\right)$; this rate matches (with respect to ϵ) the sample complexity of passive learning under the same assumptions (Massart et al., 2006) ignoring θ. Next, by rewriting Theorem 3.2 we provide a different upper bound that depends on $R(f^*)$.

Theorem 3.3. *Under the assumption that the Bernstein condition holds, for any $\delta > 0$, with probability $1-\delta$, for all $t \in [T]$, the following holds for the label requesting* *indicator P_t of IWAL-σ: $\mathbb{E}_{x\sim\mathcal{D}_x}\left[P_t \mid \mathcal{F}_{t-1}\right] \leq 2\theta r_t$, where $r_t = 2R(f^*) + \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}$ and $\tilde{\Delta}_t = \max_{f\in\mathcal{H}_t}(\hat{\Delta}_{f,f^*,t})$.*

Proof. First, we start by bounding $D(f, f^*)$, where $f \in \mathcal{H}_t$.

$$\mathbb{E}\left[|\ell(f(x), y) - \ell(f^*(x), y)|\right] \leq \mathbb{E}\left[\ell(f(x), y) + \ell(f^*(x), y)\right] = R(f) + R(f^*) \tag{3.18}$$

$$\leq 2R(f^*) + \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}, \tag{3.19}$$

where in the last inequality Theorem 3.1 is used and holds with high probability $1-\delta$. Next, we can write

$$\begin{aligned}\mathbb{E}[P_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}[\max_{f,g\in\mathcal{H}_t} \mathcal{L}(f(x), g(x)) \mid \mathcal{F}_{t-1}] \leq 2\mathbb{E}[\max_{f\in\mathcal{H}_t} \mathcal{L}(f(x), f^*(x)) \mid \mathcal{F}_{t-1}] \\ &\leq 2\mathbb{E}[\max_{f\in B(f^*,r_t)} \mathcal{L}(f(x), f^*(x)) \mid \mathcal{F}_{t-1}] \leq 2\theta r_t.\end{aligned}$$

□

Corollary 3.5. *The number of labels queried by IWAL-σ after T rounds is*

$$\sum_{t=1}^{T} \mathbb{E}[P_t \mid \mathcal{F}_{t-1}] \leq \sum_{t=1}^{T} 2\theta(2R(f^*) + \tilde{\Delta}_t + 2\tilde{\Delta}_{t-1}) \leq 2\theta(2TR(f^*) + \tilde{O}\left(T^{1-\frac{2}{4-\beta}}\right)),$$

where in the last inequality we used Corollary 3.3, then solved the integration.

Corollary 3.5 shows that the number of labels requested could be improved down to $O(\theta T^{1-\frac{2}{4-\beta}})$ when $R(f^*) \leq O(T^{-\frac{2}{4-\beta}})$. When $R(f^*)$ is adequately small, the label complexity of IWAL-σ is $\tilde{O}\left(\theta\epsilon^{-\frac{2-\beta}{2}}\right)$.

Algorithm 2: IWAL-$\sigma(\mathcal{H}, \delta, T\ )$

```
$\mathcal{H}_1 = \mathcal{H}$
for $t \in [T]$ do
    $\delta_t^\sigma = \delta_t' = \frac{\delta}{2|\mathcal{H}|^2 t(t+1)}$
    Receive $x_t$
    $P_t \leftarrow \max_{f,g \in \mathcal{H}_t} \mathcal{L}(f(x_t), g(x_t))$
    Sample $Q_t$ from $Bernoulli(P_t)$
    if $Q_t$ then  $y_t \leftarrow Label(x_t)$

    if $t \bmod 2 == 0$ then
        $\hat{f}_t \leftarrow \arg\min_{f \in \mathcal{H}_t} L_t(f)$
        for $f \in \mathcal{H}_t$ do
            $$\hat{V}_{f,\hat{f}_t,t} = \sum_{i=0}^{t/2-1} \left[\frac{Q_{2i+1}(x_{2i+1})}{p_{2i+1}(x_{2i+1})}\Big(\ell(f(x_{2i+1}), y_{2i+1}) - \ell(\hat{f}_t(x_{2i+1}), y_{2i+1})\Big) - \frac{Q_{2i+2}(x_{2i+2})}{p_{2i+2}(x_{2i+2})}\Big(\ell(f(x_{2i+2}), y_{2i+2}) - \ell(\hat{f}_t(x_{2i+2}), y_{2i+2})\Big)\right]^2$$

            $\hat{\Delta}_{f,\hat{f}_t,t} = \frac{2\sqrt{\left(\hat{V}_{f,\hat{f}_t,t} + 16\sqrt{\frac{t}{4}\ln(1/\delta_t^\sigma)}\right)\ln(4\ln(t)/\delta_t')}}{t}$
            $\mathcal{H}_{t+1} \leftarrow \{\{f\} + \mathcal{H}_{t+1} : L_t(f) \leq L_t(\hat{f}_t) + \hat{\Delta}_{f,\hat{f}_t,t}\}$
        end
    end
end
```

3.6 Special case of zero-one loss

In this section, we focus on the zero-one loss. This loss has a variety of applications since it is a natural loss to use for classification. Besides its applications, our particular interest in zero-one loss is its specific characteristics whose utilization enables us to achieve an exponential improvement over passive learning and recover known exponential improvements of label complexity achieved by Hanneke (2009); Koltchinskii (2010); Huang et al. (2015); however, we manage to achieve this result by slightly modifying an algorithm for general losses.

The first important characteristic of zero-one loss is that estimating $\sigma^2_{f,g,t}$ does not require labeled samples. It is because for zero-one loss, as long as the predictions by $f, g \in \mathcal{H}$ are not the same the loss difference will be 1; it is 0 otherwise. Therefore, we do not need the labels to estimate $\sigma^2_{f,g,t}$. Consequently, an algorithm can look at as many samples as necessary to estimate $Var(Z_{f,g,t} \mid \mathcal{F}_{t-1})$ without labeling any of the samples. To estimate $Var(Z_{f,g,t} \mid \mathcal{F}_{t-1})$ for a fixed f, g, and t we utilize Theorem 4 of Maurer and Pontil (2009) using t i.i.d. samples (which are the same first t samples given to the algorithm by Nature). However, a sample could be reused for any f, g and t, and thus, at any round t, we can use the already seen t samples to estimate $Var(Z_{f,g,t} \mid \mathcal{F}_{t-1})$ for any f and g.

The second feature of zero-one loss is that the Bernstein condition can be written in form of the Tsybakov noise condition, since for zero-one loss $(\ell(f(x), y) - \ell(g(x), y))^2 = \mathbf{1}\left[\ell(f(x), y) \neq \ell(g(x), y)\right]$.

Definition 3.3 (Tsybakov noise condition)**.** *A learning problem D with hypothesis class $\mathcal{H}$ satisfies the Tsybakov noise condition with exponent $\alpha \in [0, 1]$ and non zero constant C if*

$$P(f(X) \neq f^*(X)) \leq C(R(f) - R(f^*))^{\alpha} \qquad \textit{for all } f \in \mathcal{H}.$$

We use a different technique to estimate the variance in this section; however, in general, we use a similar analysis. Denote by $U_{f,g,t} = \mathbf{1}\left[f(x_t) \neq g(x_t)\right]$, and $\hat{U}_{f,g,T} = \frac{1}{T}\sum_{t=1}^{T} U_{f,g,t}$. Finally, let $\mu_{f,g} = \mathbb{E}[\mathbf{1}\left[f(X) \neq g(X)\right]]$. First, we estimate $\sigma^2_{f,g,T}$ by an empirical quantity denoted by $\hat{\sigma}^2_{f,g,T}$, such that with high probability, $\sigma^2_{f,g,T} \leq \hat{\sigma}^2_{f,g,T} + \epsilon$, where ϵ is the estimation error.

To estimate $\sigma^2_{f,g,T}$, first, we show that $\mathsf{Var}[Z_{f,g,t} | \mathcal{F}_{t-1}] \leq \Pr\left(f(X) \neq g(X)\right)$, using the fact that in the case of zero-one loss, $P_t \in \{0, 1\}$ and $Q_t = P_t$. We conclude that $\sigma^2_{f,g,T} \leq T \cdot \Pr\left(f(X) \neq g(X)\right)$. Therefore, to get an empirical upper bound for $\sigma^2_{f,g,T}$, it suffices to construct an estimator for $\Pr\left(f(X) \neq g(X)\right)$. Using this estimator and an empirical Bernstein bound (Maurer and Pontil, 2009), we can then get an upper confidence bound for $\sigma^2_{f,g,T}$ denoted by $\hat{\sigma}^2_{f,g,T} = T \cdot \left(\hat{U}_{f,g,T} + \sqrt{\frac{2\hat{U}_{f,g,T} \ln \frac{2}{\delta}}{T-1}} + \frac{7 \ln \frac{2}{\delta}}{3(T-1)}\right)$, where δ is the confidence variable. Since $\sigma^2_{f,g,T} \leq \hat{\sigma}^2_{f,g,T}$, we can use $\hat{\sigma}^2_{f,g,T}$ in our algorithm.

Lemma 3.5. *We have $\sigma^2_{f,g,T} \leq T \cdot \Pr\left(f(X) \neq g(X)\right)$.*

Proof. For the sake of readability, we introduce the more compact notation

$$\ell_{f-g}(X_t, Y_t) := \ell(f(X_t), Y_t) - \ell(g(X_t), Y_t)$$

and $R_{f-g} := R(f) - R(g)$. Recall that

$$Z_{f,g,t} = \mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \left(\frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t) - R_{f-g}\right).$$

First, observe that

$$\begin{aligned}
\mathsf{Var}[Z_{f,g,t} \mid \mathcal{F}_{t-1}] &\leq \mathbb{E}\left[Z^2_{f,g,t} \mid \mathcal{F}_{t-1}]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \left(\frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t) - R_{f-g}\right)^2 \;\middle|\; \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \left(\frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t)^2 - 2\frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t)R_{f-g} + R^2_{f-g}\right) \;\middle|\; \mathcal{F}_{t-1}\right],
\end{aligned}$$

where the last equality uses the fact that $Q_t^2 = Q_t$, that $P_t^2 = P_t$ (since $P_t \in \{0, 1\}$ for zero-one loss). Next, for zero-one loss, whenever $f, g \in \mathcal{H}_t$, we have from the definition of P_t that $Q_t = P_t$ and $P_t \geq \ell_{f-g}(X_t, Y_t)$. Therefore, the last line above is equal to

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \left(\frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t)^2 - R^2_{f-g}\right) \;\middle|\; \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \frac{Q_t}{P_t}\ell_{f-g}(X_t, Y_t)^2 \;\middle|\; \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \cdot \frac{Q_t}{P_t}\left|f(X_t) - g(X_t)\right| \;\middle|\; \mathcal{F}_{t-1}\right].
\end{aligned}$$

Finally, again using $Q_t = P_t \geq \ell_{f-g}(X_t, Y_t)$ when $f, g \in \mathcal{H}_t$, the last line above is equal to

$$\mathbf{1}\left[f, g \in \mathcal{H}_t\right] \Pr\left(f(X_t) \neq g(X_t)\right) \leq \Pr\left(f(X_t) \neq g(X_t)\right).$$

In conclusion, we have shown that

$$\mathsf{Var}[Z_{f,g,t} \mid \mathcal{F}_{t-1}] \leq \Pr\left(f(X_t) \neq g(X_t)\right).$$

Now, we look at getting an empirical version of $\sigma^2_{f,g,T}$. We have

$$\begin{aligned}\sigma^2_{f,g,T} &= \sum_{t=1}^{T} \mathsf{Var}[Z_{f,g,t} \mid \mathcal{F}_{t-1}] \\ &\leq \sum_{t=1}^{T} \Pr\left(f(X_t) \neq g(X_t)\right) \\ &= T \cdot \Pr\left(f(X) \neq g(X)\right),\end{aligned}$$

where $X \sim D$. □

Lemma 3.6. *For any $f, g \in \mathcal{H}$, with probability at least $1-\delta$, we have $\sigma^2_{f,g,T} \leq \hat{\sigma}^2_{f,g,T}$, where $\hat{\sigma}^2_{f,g,T} = T \cdot \left(\hat{U}_{f,g,T} + \sqrt{\frac{2\hat{U}_{f,g,T}\ln\frac{2}{\delta}}{T-1}} + \frac{7\ln\frac{2}{\delta}}{3(T-1)}\right)$.*

Proof. From Theorem 4 of Maurer and Pontil (2009), we know if $W_1, \ldots, W_n$ are i.i.d. Bernoulli samples with true mean μ, with high probability at least $1-\delta$,

$$\mu \leq \hat{W}_n + \sqrt{\frac{2V_n(\mathbf{W})\ln\frac{2}{\delta}}{n}} + \frac{7\ln\frac{2}{\delta}}{3(n-1)},$$

where $V_n(\mathbf{W}) := \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}(W_i - W_j)^2$ is the sample variance of $\mathbf{W} = (W_1, \ldots, W_n)$ and $\hat{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i$. Also, since W_i is $\{0,1\}$-valued sample drawn from a Bernoulli distribution, and $0 \leq \hat{W}_n \leq 1$, we have $V_n(\mathbf{W}) = \frac{n}{n-1}\hat{W}_n(1-\hat{W}_n) \leq \frac{n}{n-1}\hat{W}_n$. Therefore, also with probability at least $1-\delta$,

$$\mu \leq \hat{W}_n + \sqrt{\frac{2\hat{W}_n\ln\frac{2}{\delta}}{n-1}} + \frac{7\ln\frac{2}{\delta}}{3(n-1)}. \tag{3.20}$$

Since $U_{f,g,1}, \ldots, U_{f,g,T}$ are $\{0,1\}$-valued, i.i.d. samples drawn from a Bernoulli distribution with true mean $\mu_{f,g}$, we can use (3.20), to estimate $\mu_{f,g}$. Also, from Lemma 3.5, we know that $\sigma^2_{f,g,T} \leq T \cdot \mu_{f,g}$. Thus, we can write

$$\sigma^2_{f,g,T} \leq T \cdot \left(\hat{U}_{f,g,T} + \sqrt{\frac{2\hat{U}_{f,g,T}\ln\frac{2}{\delta}}{T-1}} + \frac{7\ln\frac{2}{\delta}}{3(T-1)}\right).$$

□

To bound the generalization error, we bound $\tilde{\Delta}_T$, in order to do that, we have to bound $\hat{U}_{f,g,T}$ first. Using Bennett's inequality, we prove that $\hat{U}_{f,g,T} \leq \mu_{f,g}+\sqrt{\frac{2\mu_{f,g}\ln\frac{1}{\delta}}{T}}+\frac{\ln\frac{1}{\delta}}{3T}$. Next, using the upper bound on $\hat{U}_{f,g,T}$, we can bound $\hat{\Delta}_{f,g,T}$ in terms of $\mu_{f,g}$ and T. Finally, we can upper bound the generalization error.

Lemma 3.7. *For any $t > 0$ and a fixed $f, g \in \mathcal{H}$, with high probability at least $1-\delta$,*

$$\hat{U}_{f,g,T} \leq \mu_{f,g} + \sqrt{\frac{2\mu_{f,g}\ln\frac{1}{\delta}}{T}} + \frac{\ln\frac{1}{\delta}}{3T}. \tag{3.21}$$

Proof. Let $W_{f,g,t} = 1 - U_{f,g,t}$. Also, let $U_{f,g}$ be an independent copy of $U_{f,g,1}$ and set $W_{f,g} = 1 - U_{f,g}$. Using Bennett's inequality (See Theorem 3 of Maurer and Pontil (2009)), with high probability at least $1-\delta$, we can write

$$\mathbb{E}[W_{f,g}] - \frac{1}{T}\sum_{t=1}^{T} W_{f,g,t} \leq \sqrt{\frac{2\,\mathsf{Var}(W_{f,g})\ln\frac{1}{\delta}}{T}} + \frac{\ln\frac{1}{\delta}}{3T}.$$

Since $W_{f,g,t} = 1 - Z_{f,g,t}$ and $\mathsf{Var}(W_{f,g}) = \mathsf{Var}(1-U_{f,g}) = \mathsf{Var}(U_{f,g})$, we can write

$$1 - \mathbb{E}[U_{f,g}] - (1 - \frac{1}{T}\sum_{t=1}^{T} U_{f,g,t}) \leq \sqrt{\frac{2\,\mathsf{Var}(U_{f,g})\ln\frac{1}{\delta}}{T}} + \frac{\ln\frac{1}{\delta}}{3T}.$$

Next, since $\mathsf{Var}(U_{f,g}) = \mu_{f,g}(1-\mu_{f,g}) \leq \mu_{f,g}$, we get

$$\hat{U}_{f,g,T} \leq \mu_{f,g} + \sqrt{\frac{2\mu_{f,g}\ln\frac{1}{\delta}}{T}} + \frac{\ln\frac{1}{\delta}}{3T}.$$

□

Lemma 3.8. *For any $f, g \in \mathcal{H}_T$ with high probability $1-\delta$, we have* $\hat{\Delta}_{f,g,T} \leq \max\left\{\sqrt{\frac{\mu_{f,g}}{T}}, \frac{\mu_{f,g}^{1/4}}{T^{3/4}}, \frac{1}{T}, \frac{\mu_{f,g}^{1/8}}{T^{7/8}}\right\}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}$.

Proof. For the sake of readability, we summarize $\hat{\sigma}_{f,g,T}, \hat{U}_{f,g,T}$, and $\mu_{f,g}$ by $\hat{\sigma}_T, \hat{U}_T$, and

μ respectively. We can write,

$$\begin{aligned}
\hat{\Delta}_{f,g,T} &= \frac{\sqrt{\hat{\sigma}_T^2 \ln(\frac{4\ln(T)}{\delta'_T})}}{T} = \frac{\sqrt{T\left(\hat{U}_T + \sqrt{\frac{\hat{U}_T}{T}} + \frac{1}{T}\right)\ln(\frac{4\ln(T)}{\delta'_T})}}{T} \\
&\leq \frac{\sqrt{\left(\hat{U}_T + \sqrt{\frac{\hat{U}_T}{T}} + \frac{1}{T}\right)\ln(\frac{4\ln(T)}{\delta'_T})}}{\sqrt{T}} \\
&\leq \frac{\sqrt{\left(\mu + \sqrt{\frac{\mu}{T}} + \frac{1}{T} + \sqrt{\frac{\mu+\sqrt{\frac{\mu}{T}}+\frac{1}{T}}{T}} + \frac{1}{T}\right)\ln(\frac{4\ln(T)}{\delta'_T})}}{\sqrt{T}} \\
&\leq \max\left\{\sqrt{\frac{\mu}{T}},\ \frac{\mu^{1/4}}{T^{3/4}},\ \frac{1}{T},\ \frac{\mu^{1/8}}{T^{7/8}}\right\}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}
\end{aligned}$$

□

Lemma 3.9. *The generalization error is* $R(\hat{f}_T) - R(f^*) \leq O\left(\left(\frac{C}{T}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}\right)^{\frac{1}{2-\alpha}}\right)$.

Proof. First, observe that since the Tsybakov noise condition holds, for any $f \in \mathcal{H}_T$,

$$\mu_{f,f^*} = Pr(f(X) \neq f^*(X)) \leq C(R(f) - R(f^*))^{\alpha}, \tag{3.22}$$

and from Theorem 3.1, we know $R(\hat{f}_T) - R(f^*) \leq \hat{\Delta}_{\hat{f}_T,f^*,T}$. By combining (3.22), Lemma 3.8, and $R(\hat{f}_T) - R(f^*) \leq \hat{\Delta}_{\hat{f}_T,f^*,T}$ we can write

$$\mathcal{R}_T \leq \max\left\{\sqrt{\frac{C\mathcal{R}_T^{\alpha}}{T}},\ \frac{C^{1/4}\mathcal{R}_T^{\alpha/4}}{T^{3/4}},\ \frac{1}{T},\ \frac{C^{1/8}\mathcal{R}_T^{\alpha/8}}{T^{7/8}}\right\}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})} \tag{3.23}$$

where $\mathcal{R}_T = R(\hat{f}_T) - R(f^*)$. Solving (3.23) will give us the desired result. □

Label Complexity under zero-one loss

To find the number of queried samples for IWAL-σ in case of zero-one loss, we use an analysis similar to the Section 3.4.

Theorem 3.4. *Under the assumption that the Tsybakov condition holds, for any* $\delta > 0$, *with probability* $1 - \delta$, *for all* $t \in [T]$, *the following holds for the label requesting* *indicator P_t of IWAL-σ: $\mathbb{E}_{x\sim\mathcal{D}_x}[P_t \mid \mathcal{F}_{t-1}] \leq 2\theta r_t$, where $\tilde{\Delta}_t = \max_{f\in\mathcal{H}_t}(\hat{\Delta}_{f,f^*,t}))$ and $r_t = C(\tilde{\Delta}_t + 2\tilde{\Delta}_{t-1})^\alpha$*

Proof. The proof of Theorem 3.4 is similar to proof of Theorem 3.2, except we bound $D(f, f^*)$ differently. The following is the new bound for $D(f, f^*)$. We start by definition of $D(f, f^*)$,

$$\mathbb{E}\left[|\ell(f(x), y) - \ell(f^*(x), y)|\right] \leq C(R(f) - R(f^*))^\alpha \leq C(\tilde{\Delta}_t + 2\tilde{\Delta}_{t-1})^\alpha \tag{3.24}$$

where in the first inequality we used the Tsybakov condition and then, Theorem 3.1 is used. □

Next, first, we upper bound $\tilde{\Delta}_t$ and then, we use this upper bound to upper bound the number of samples queried by the algorithm.

Corollary 3.6. $\tilde{\Delta}_T \leq \left(\frac{C}{T}\right)^{\frac{1}{2-\alpha}} \left(\sqrt{\ln\left(\frac{4\ln(T)}{\delta'_T}\right)}\right)^{\frac{\alpha}{4-2\alpha}}$ *where* $\tilde{\Delta}_T = \max_{f\in\mathcal{H}_T}(\hat{\Delta}_{f,f^*,T}))$.

Proof. Using an argument similar to Lemma 3.9, we first show

$$R(f) - R(f^*) \leq O\left(\frac{C}{T}\ln\left(\frac{4\ln(T)}{\delta'_T}\right)\right)^{\frac{1}{2-\alpha}} \tag{3.25}$$

for any $f \in \mathcal{H}_T$ with high probability $1-\delta$. In the argument, we use $R(f) - R(f^*) \leq \tilde{\Delta}_T + 2\tilde{\Delta}_{T-1}$ instead of $R(\hat{f}_T) - R(f^*) \leq \tilde{\Delta}_T$, both from Theorem 3.1. Then, we upper bound $\hat{\Delta}_{f,f^*,T}$ using Lemma 3.8 which gives us,

$$\hat{\Delta}_{f,f^*,T} \leq \max\left\{\sqrt{\frac{\mu_{f,f^*}}{T}}, \frac{\mu_{f,f^*}^{1/4}}{T^{3/4}}, \frac{1}{T}, \frac{\mu_{f,f^*}^{1/8}}{T^{7/8}}\right\}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}. \tag{3.26}$$

Next, since $\mu_{f,f^*} \leq C(R(f) - R(f^*))^\alpha \leq C\left(3\tilde{\Delta}_T\right)^\alpha$, where the first inequality holds because of the Tsybakov noise condition. For any $f \in \mathcal{H}_T$, we conclude that $\hat{\Delta}_{f,f^*,T} \leq \left(\frac{C}{T}\right)^{\frac{1}{2-\alpha}} \left(\sqrt{\ln\left(\frac{4\ln(T)}{\delta'_T}\right)}\right)^{\frac{\alpha}{4-2\alpha}}$ thus, $\tilde{\Delta}_T \leq \left(\frac{C}{T}\right)^{\frac{1}{2-\alpha}} \left(\sqrt{\ln\left(\frac{4\ln(T)}{\delta'_T}\right)}\right)^{\frac{\alpha}{4-2\alpha}}$. □

Corollary 3.7. *The number of labels queried by IWAL-σ after T rounds is $\sum_{t=1}^{T}\mathbb{E}[P_t \mid \mathcal{F}_{t-1}] \leq \tilde{O}\left(\theta C^2 T^{\frac{2-2\alpha}{2-\alpha}}\right)$ for $\alpha < 1$, and for $\alpha = 1$, $\sum_{t=1}^{T}\mathbb{E}[P_t \mid \mathcal{F}_{t-1}] \leq \tilde{O}\left(\theta C^2\log(T)\right)$.*

Proof. The proof is similar to proof of Corollary 3.4 except that instead of Corollary 3.3, we use Corollary 3.6. □

With a closer look at Corollary 3.7 and Lemma 3.9, we can see if $\alpha < 1$, we obtain a polynomial label complexity of $O(\theta\epsilon^{2\alpha-2})$, and when $\alpha = 1$ we achieve a logarithmic label complexity.

3.7 Future Work

One of the disadvantages of IWAL is that it uses an effective version space, which in many cases, is not efficiently implementable. Works like those of Beygelzimer et al. (2010) and Huang et al. (2015) proposed a solution that uses the idea of importance weighted sampling while simultaneously avoiding the use of an effective version space. Their suggested solutions are efficient under the assumption of an ERM oracle and are specifically designed for zero-one loss. As future work, we would like to investigate, under convex losses and for particular hypotheses classes like linear separators, whether IWAL-σ or a suitable variant affords an efficient implementation.

Chapter 4

Experiments

In this chapter, we aim to analyze the performance of IWAL-σ on synthetic data. We also compare IWAL-σ's performance to IWAL and passive learning's performance. First, we discuss our setup for the experiments. Second, we explain how and why IWAL-σ has been modified for implementations. After that, we see our results for different loss functions and the amount of noise.

4.1 Experiment Setup

During this chapter, we refer to the number of dimensions of our samples by d. The number of dimensions for each experiment is mentioned before the plots.

Generating data. The synthetic data is created by uniformly random sampling 55000 points over a d-ball using Muller's technique, of which 5000 samples will be used for training and the rest for evaluation. To label our samples, for each sample x, we create a tuple (x, y), where $y = f_v(x) := \operatorname{sgn}(x \cdot v)$. Vector v of dimension d corresponds the perfect classifier before adding any noise and it is defined as $v = \frac{1}{\sqrt{d}}\mathbf{1}$. Vector v is associated with function f_v and is also added to the hypothesis set. Moreover, we add noise to the samples by flipping a coin for each sample. The probability of success for each coin is determined based on the amount of noise we would like to add. We do our experiments for different noise values. Note that even if we do not add noise, f_v might not be the best hypothesis (the hypothesis with lowest test error on our dataset) for some loss functions like logistic loss. The only case that f_v is guaranteed to be the best hypothesis is under zero-one loss with no noise, known as the *realizable case*.

Generating hypothesis set. We create a set of randomly drawn unit homogeneous linear separators. To generate a random hypothesis, we first draw a vector from a d-dimensional normal distribution and normalize it to be a unit vector. The cardinality of our initial hypothesis set is 5000. To make the problem even harder, we find the best hypothesis f on the training data. Next, we create $50 \cdot d$ new hypotheses that are similar to f. To create a new hypothesis similar to f, I) we uniformly draw a sample i, between 1 and d. II) we draw a sample s from a zero mean Gaussian distribution with standard deviation 0.2. A new hypothesis g is created by adding s to the i^{th} index of f.

Implementations. We have implemented IWAL precisely as it appears in the original work (Beygelzimer et al., 2009). IWAL-σ has been implemented similar to the Algorithm 2 but, there is one modification to the upper deviation bound. The upper deviation bound has been modified to adapt to the loss range. In Chapter 3, we assumed that the loss is bounded by 1. However, this assumption is not valid for the experiments. For this reason, we modify our theoretical results to adapt to the loss range denoted by b. The only actual difference in the algorithm is in $\hat{\Delta}_{f,g,t}$, where $\hat{\Delta}_{f,g,T}$ new definition is

$$\hat{\Delta}_{f,g,T} = \frac{1}{T}\max\left(2\sqrt{\hat{V}_{f,g,T} + 16\sqrt{\frac{T}{4}}\ln(2/\delta_T^{\sigma})}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}, 6b\ln(\frac{4\ln(T)}{\delta'_T})\right).$$

For passive learning, we use empirical risk minimization. It is important to remember that any passive learning algorithm visits only the first K samples provided by Nature, where K is the maximum number of labels queried by IWAL and IWAL-σ. On the other hand, an active learning algorithm (IWAL or IWAL-σ in our case) that has labeled K samples potentially has seen more samples. In other words, the labeled samples visited by IWAL, IWAL-σ, and ERM can be different. Active learning algorithms will stop learning once there is only one hypothesis left in the effective version space or there is no training data left.

Experiments have been done under a specific loss function for a certain amount of noise for all three algorithms. Each experiment is repeated 50 times, and the results are averaged; we refer to each of these repetitions by a trial. Two sets of plots are given for each experiment.

The first set of plots answers the question of *what is the excess risk of* $\hat{f}_t$ *after* *querying i labels by each algorithm?*. The second set of plots approaches the question of *what is the excess risk of the worst f in the effective version space after querying i labels by each algorithm?* The reason we care about this question is that from a theoretical standpoint, there is no guarantee that $\hat{f}_t$ is significantly better than any other hypothesis left in the effective version space. We can argue that comparing the worst function in the effective version space is the right way of comparing two active learning algorithms based on the idea of effective version space. This question does not apply to passive learning algorithms since ERM does not maintain a effective version space. For each question, there exists two series of plots. The first series depicts excess risk only. The second series depicts the excess risk in a log scale plot with a confidence band. The confidence band is determined by the standard deviation of excess risks over all the trials. For example, in the log scale plot under zero-one loss for $\hat{f}_t$, when the number of labels queried (NLQ) is 10, the value of the plot itself is the log of the average of excess risk of $\hat{f}_t$ after querying the tenth label, and the confidence band is the standard deviation of these values.

It is clear that not all the trials will use the same number of labels since the samples are generated randomly. In such situations, the standard deviation and the average are taken over the trials that request at least that many labels. For example, assume 30 trials request 50 labels, and the other 20 trials query 80 labels. Then, to find the confidence band for NLQ = 75, we find the standard deviation of the excess risk of those trials that query at least 75 labels.

After running the experiments by looking at Figures 4.1, 4.4, 4.7, and 4.8, we notice that IWAL-σ is not learning as quickly as IWAL. Taking a close look at the algorithm, we can see this happens mainly because the second term in $\hat{\Delta}_{f,\hat{f}_T,T}$ can be too large. The term $6b\ln(\frac{4\ln(T)}{\delta'_T})$ can be larger than 100, which to us, seems unnecessarily large in practice. To test our hypothesis, we modified the algorithm and ran another set of experiments. In this modification, the $\hat{\Delta}_{f,\hat{f}_T,T}$ is set to only the first term $2\sqrt{\hat{V}_{f,g,T} + 16\sqrt{\frac{T}{4}\ln(2/\delta^{\sigma}_T)}}\sqrt{\ln(\frac{4\ln(T)}{\delta'_T})}$, and we have altered the shrinking process to the following

$$\mathcal{H}_{t+1} \leftarrow \{f \in \mathcal{H}_t : L_t(f) \leq L_t(\hat{f}_t) + \hat{\Delta}_{f,\hat{f}_t,t} + C_0/t\},$$

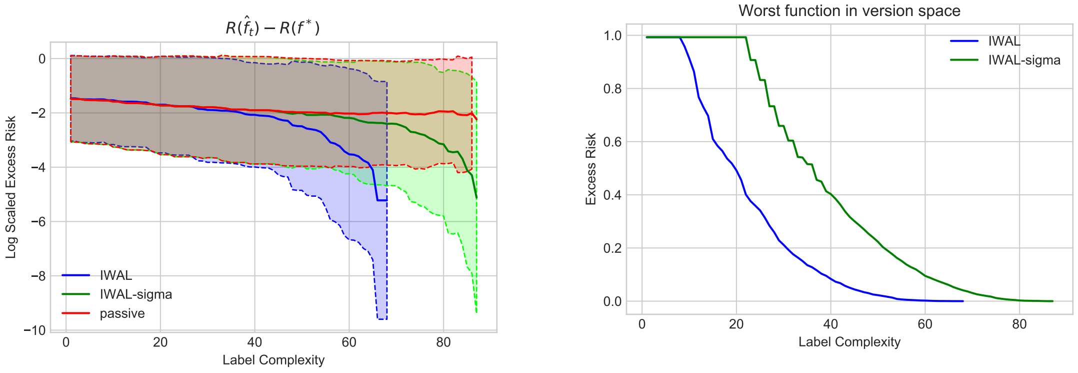
where C_0 is a constant. A good value for C_0 varies for each loss. Intuitively, the purpose of C_0 is to make sure that $L_t(f)$, $L_t(\hat{f}_t)$, $\hat{\Delta}_{f,\hat{f}_t,t}$ are large enough. This is important because if these values are too small, they might not be accurate yet.

4.2 Results and Discussion

We categorize our results based on the loss functions. All the experiments are done with $d = 3$.

4.2.1 Zero-one Loss

Zero-one loss is the most intuitive loss when it comes to classification. Interestingly, in our modification of IWAL-σ we can set $C_0 = 0$. First, we look at IWAL-σ and the other two algorithms under 5% noise. We observe that IWAL-σ converges more slowly than IWAL.



(a) Excess risk

(b) Excess risk of the worst function left in version space

Figure 4.1: Results for zero-one loss with no modification on IWAL-σ

Next, we look at how well our modified algorithm and the other two algorithms do under different amounts of noise.

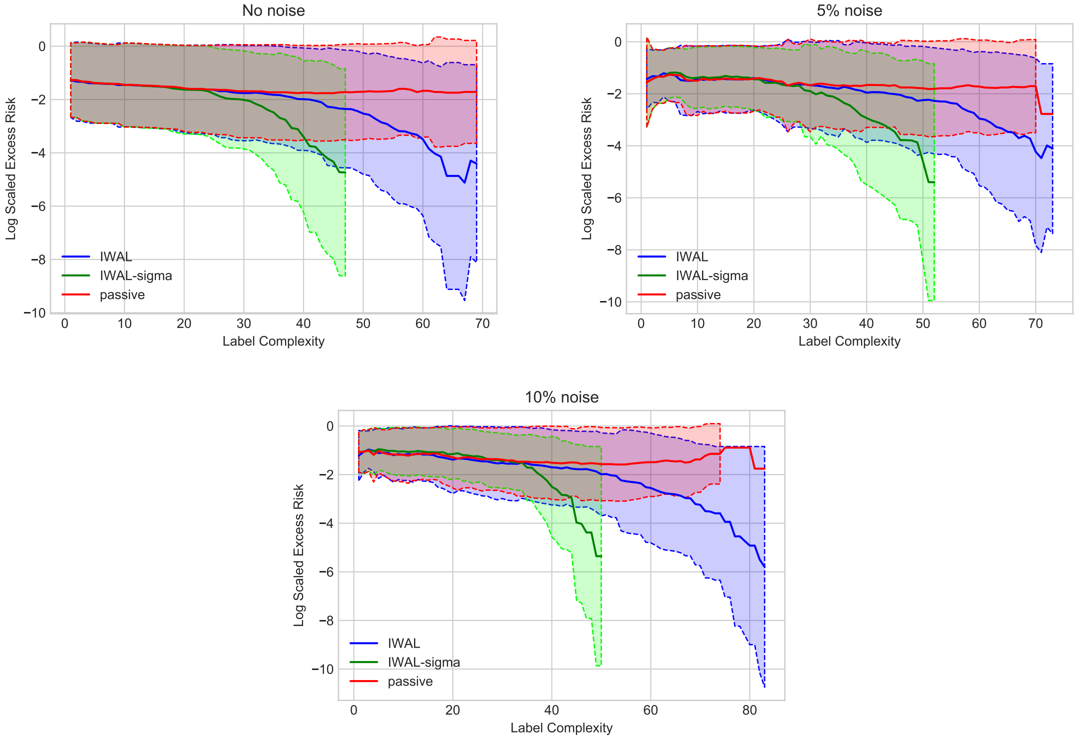


Figure 4.2: Results for zero-one loss for modified IWAL-σ

In Figure 4.2, the darker lines are the average log scaled excess risk of the best empirical hypothesis after observing t labels and the shaded area shows the confidence interval. We can see that all three algorithms are robust to noise; however, the difference is that ERM or the passive learning algorithm barely learns anything. This is not surprising since passive learners do not actively choose informative samples. An interesting observation is the small fluctuations in IWAL and IWAL-σ. These small fluctuations increase with noise.

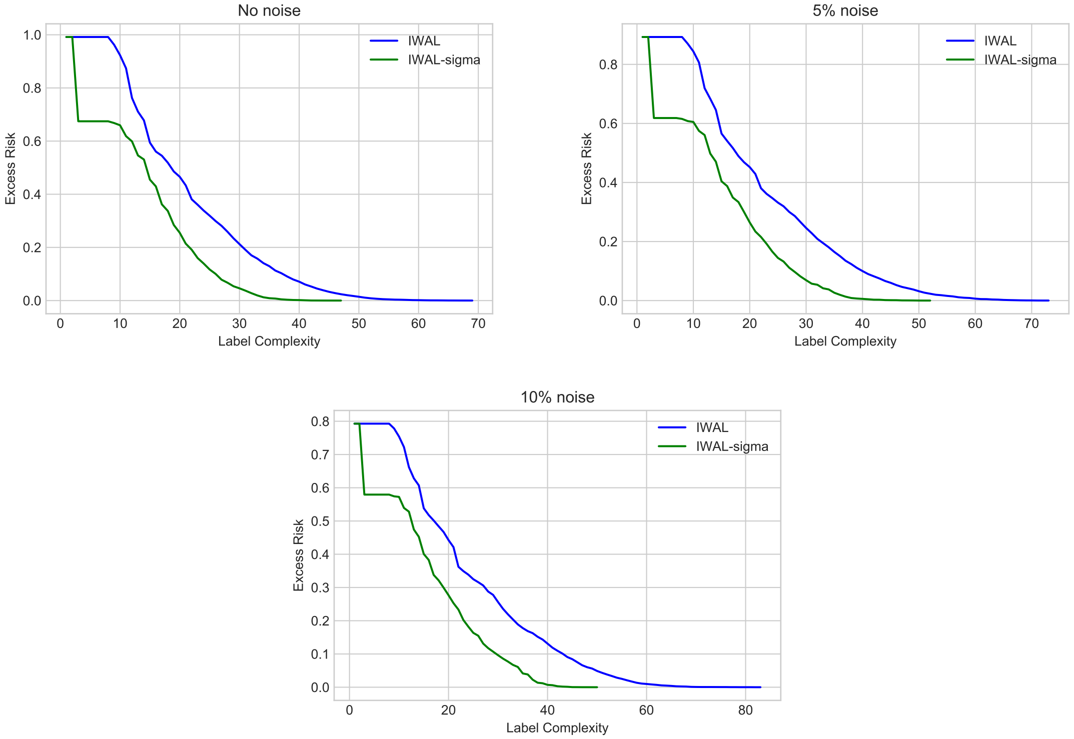
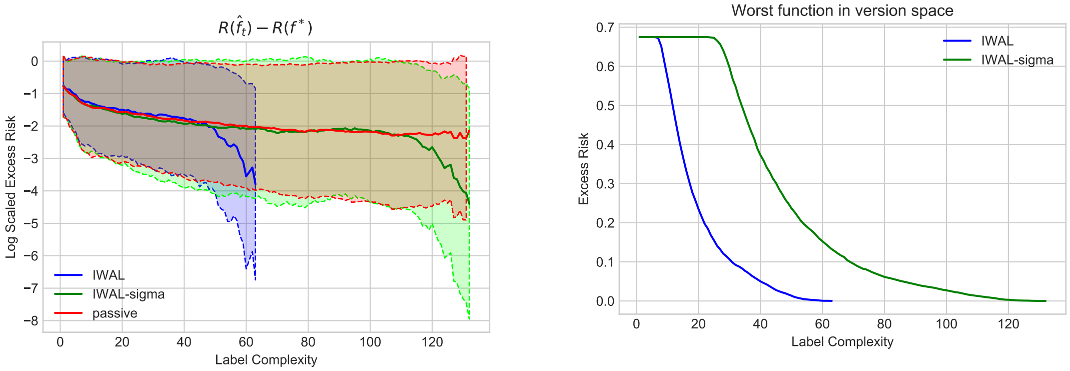


Figure 4.3: Excess risk of the worst function left in version space for modified IWAL-σ

Figure 4.3 is useful in response to two questions. First, *what is the excess risk of worst function in the effective version space after querying i labels by each algorithm?* Second, it answers the question of if the algorithms are stable, i.e., whether noise has any effect on stability of the algorithms. We can see that the graphs in Figure 4.3 are smooth despite the amount of noise. This shows that IWAL and IWAL-σ are similar in how they shrink the effective version space. As expected, the algorithms finish faster with a lower amount of noise.

4.2.2 Hinge loss

Hinge loss is a popular loss when it comes to classification. It is defined as $\ell(f(x), y) = \max(0, 1 - y \cdot f(x))$, and is commonly used for maximum margin classification. In this set of experiments, the samples' dimension is 3, and all the experiments are done by adding 5% noise. The figures below show that similar to before, the unmodified version of IWAL-σ is not better than IWAL. However, the modified version can converge faster than IWAL.



(a) Excess risk

(b) Excess risk of the worst function left in version space

Figure 4.4: Results for hinge loss with no modification on IWAL-σ

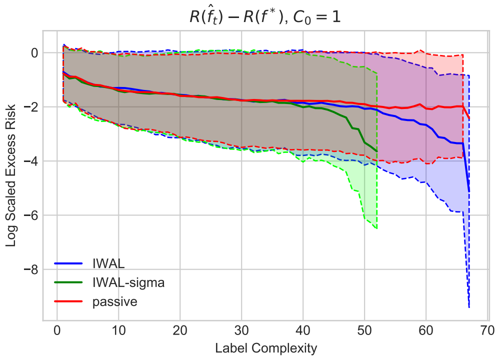


Figure 4.5: Results for hinge loss for modified IWAL-σ - 5% noise

In the Figure 4.5, where $C_0 = 1$, we can see that IWAL and IWAL-σ both are smooth and converge faster than passive learning. The only exception is in the right plot, wherein the right tail, IWAL has a significant drop. This is not because IWAL does a better job after many rounds, but it happens because only a few trials request 65 labels or more. Therefore, if all the trials would request the same amount of labels, we would not see such a drop. Next, we study the importance of C_0.

By studying the plots of the worst function in the version space, we can see the effect of C_0. The larger C_0 is, the slower IWAL-σ will shrink the effective version space. Therefore to benefit from IWAL-σ, it is vital to pick a good value for C_0. If C_0 is too large, IWAL-σ can be even slower than IWAL, and if C_0 is too small, it will shrink the version space carelessly and eliminate hypotheses with low excess risk.

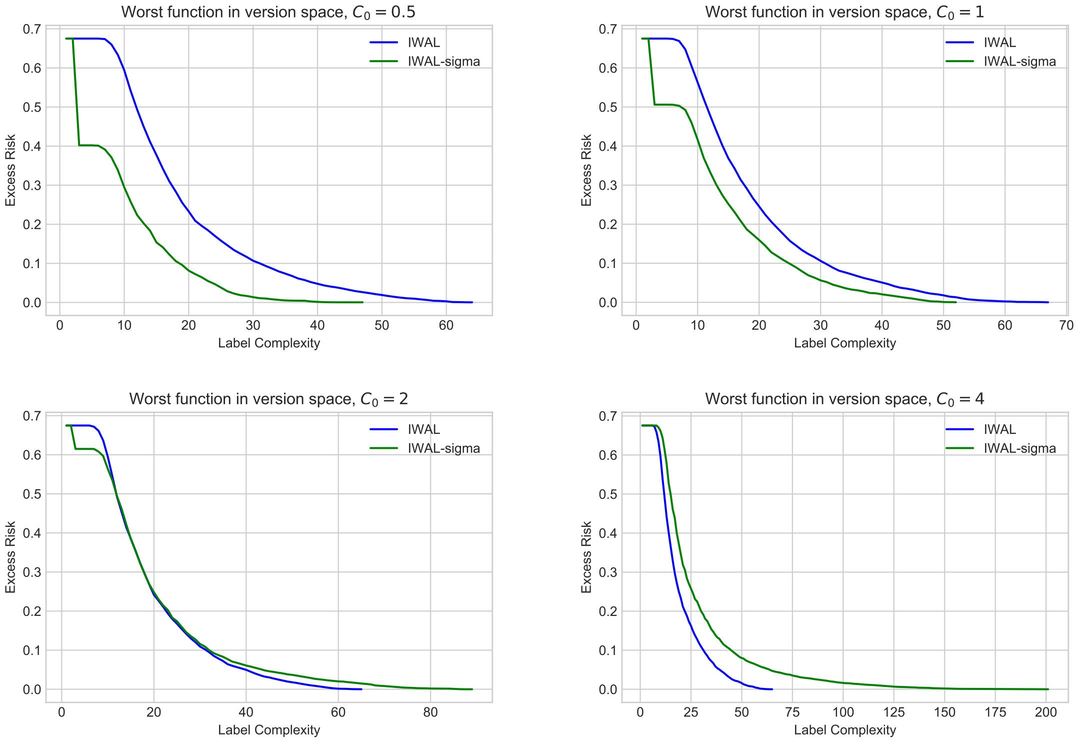


Figure 4.6: Excess risk of the worst function left in the effective version space for modified version of IWAL-σ - Hinge loss - 5% noise

4.2.3 Logistic and Squared Loss

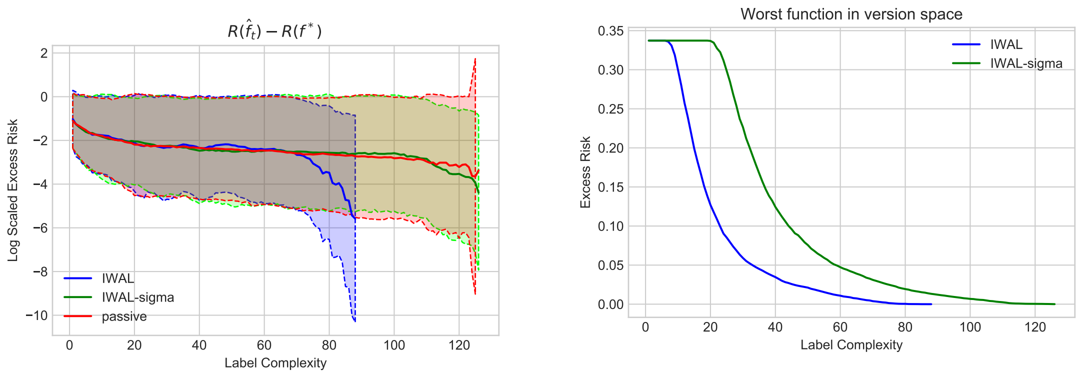
Logistic loss is of interest since it is convex and not sensitive to outliers. Logistic loss is defined as

$$\ell(f(x), y) = \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{if } y = 0, \end{cases}$$

and popular in regression problems.

Similar to logistic loss, squared loss is convex and a commonly used loss function for regression problems, however squared loss is sensitive to outliers. Squared loss is defined as $\ell(f(x), y) = (f(x) - y)^2$.

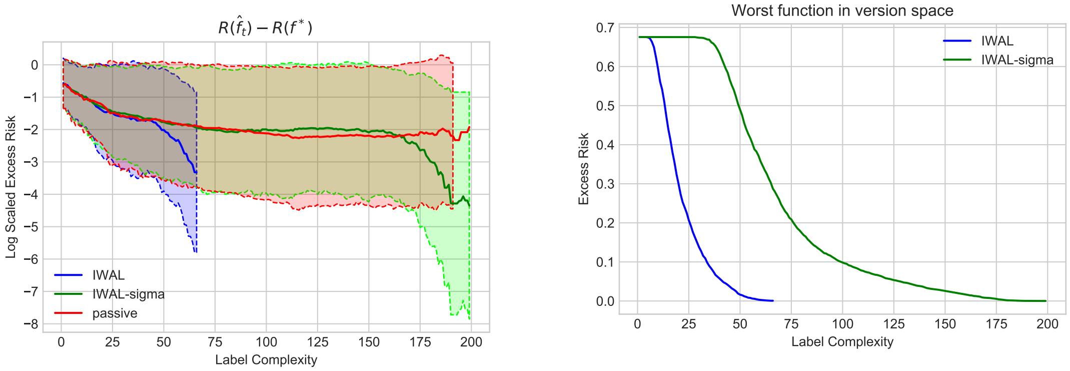
In the case of both of these losses, the results for the unmodified IWAL-σ are not desirable.



(a) Excess risk

(b) Excess risk of the worst function left in version space

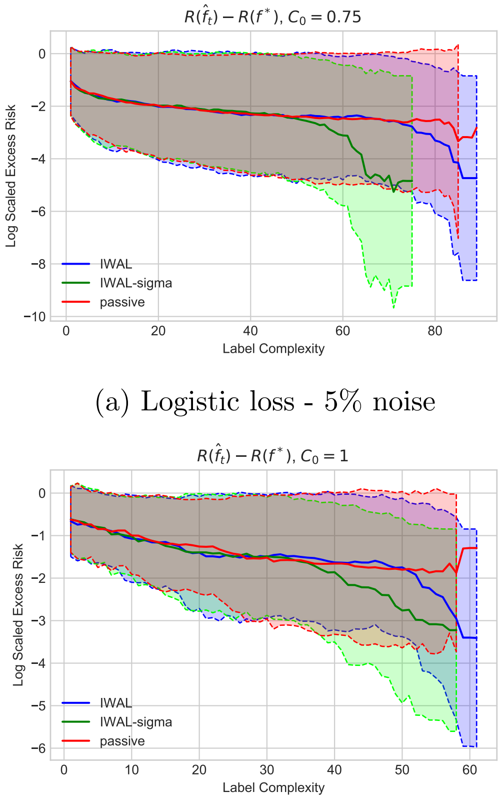
Figure 4.7: Results for logistic loss with no modification on IWAL-σ



(a) Excess risk

(b) Excess risk of the worst function left in version space

Figure 4.8: Results for squared loss with no modification on IWAL-σ



(a) Logistic loss - 5% noise

(b) Squared loss - 5% noise

Figure 4.9: Results for the modified version of IWAL-σ

In Figures 4.9a and 4.9b we can see that IWAL-σ for logistic loss and squared

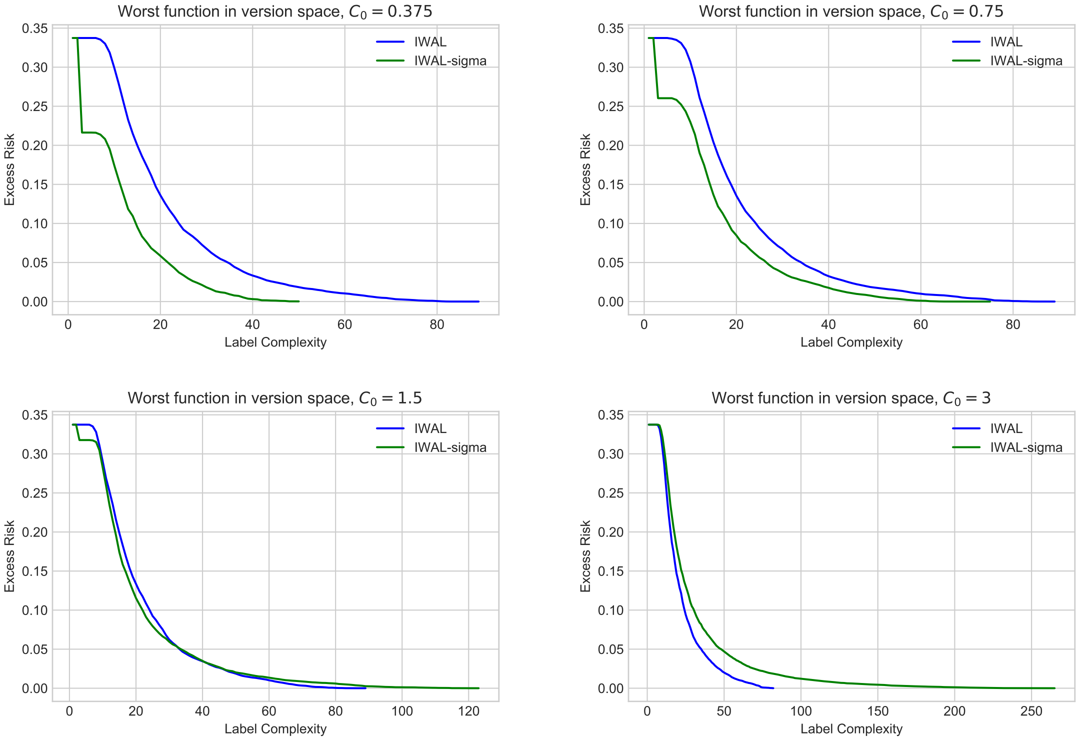


Figure 4.10: Excess risk of the worst function left in the effective version space for modified IWAL-σ - Logistic loss - 5% noise

loss is faster than IWAL and passive learning on average. It is worth mentioning that even though on average IWAL-σ looks to be faster than others, after taking the confidence intervals into account, it might not be clear that IWAL-σ is always the best. Additionally, since $R(f^*)$ is not small enough in both of these cases, at least in our experiments, significant improvement might not be possible. IWAL-σ seems to be unstable in Figure 4.9a, especially at its final rounds of execution; however Figure 4.10 shows the opposite, and is a better indicator of how stable the algorithm is. Similar to the results for hinge loss, we can see it is important to choose the value of C_0 carefully for both squared and logistic loss. If the value of C_0 is too small, the drop at the beginning of plots in Figures 4.10 and 4.11 can be too large, which could lead to the elimination of the best hypothesis or hypotheses close to the best hypothesis by IWAL-σ.

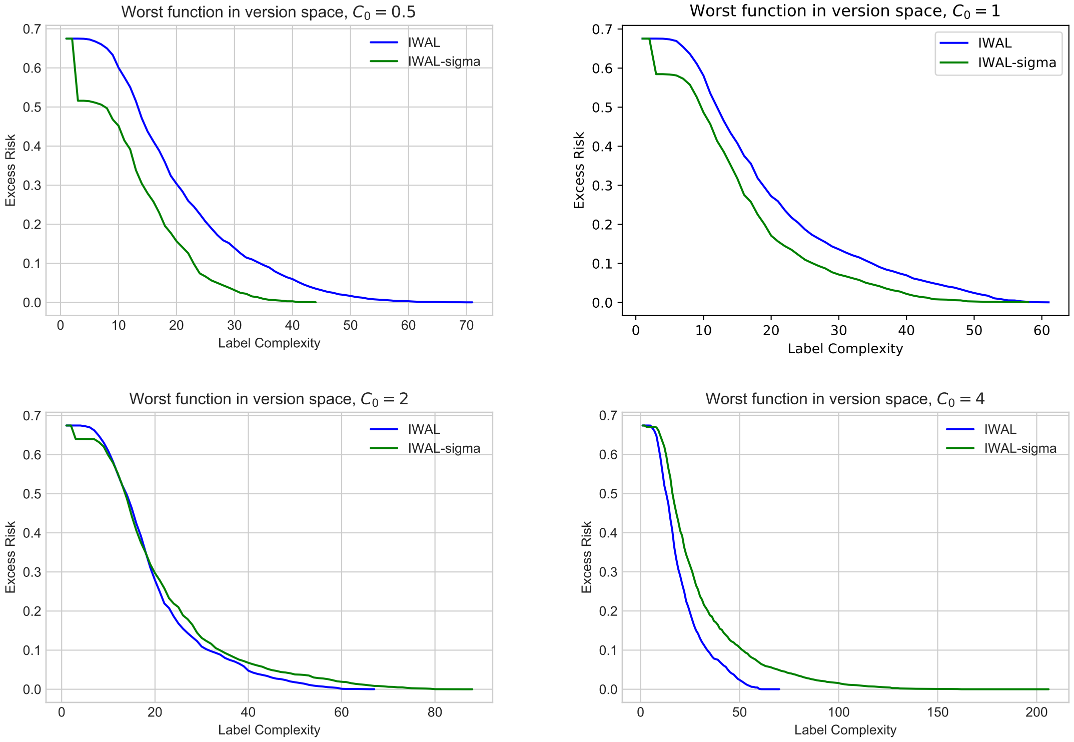


Figure 4.11: Excess risk of the worst function left in the effective version space for the modified version of IWAL-σ - Squared loss - 5% noise

4.3 Conclusion

We can see that the modified version of IWAL-σ outperforms IWAL and passive learning in some of the experiments on average. However, in some cases like squared and logistic loss, the the results do not confirm a significant improvement over passive learning. However, to fully benefit from IWAL-σ, we have to tune its only parameter C_0. An easy rule of thumb for tuning C_0 is to set it to half of the loss range. For example, for hinge loss, $C_0 = 1$ achieves acceptable performance. Of course, the biggest problem with IWAL-σ and similarly IWAL is that these algorithms are computationally tractable.

Chapter 5

Conclusion and Future Work

Active learning could be the solution to learning problems with a large amount of data, whether these data are labeled or unlabeled. For unlabeled datasets, active learning will can reduce the cost of labeling the dataset, and in the case that the data is already labeled, active learning can reduce the computational cost of learning since, in many cases it learns faster than passive learners.

There have been many attempts to propose efficient active learning algorithms in the past years. In some settings, efficient and practical algorithms have been successfully suggested. However, even efficient algorithms tend to query many labels. The number of labels that algorithms based on disagreement need to query from an oracle usually differs from the label complexity's lower bounds by at least a factor of disagreement coefficient. One of the challenges in active learning is proposing an algorithm that the number of labels it needs is closer to the lower bounds, or in other words, it does not depend on the disagreement coefficient.

Another idea that could be worked on in online active learning is to propose a new way of deciding whether to query a label or not. The work of Tosh and Dasgupta (2017) could inspire a new technique in online active learning. It might be possible to focus on the average disagreement on a sample instead of focusing on the maximum disagreement in a hypothesis set.

In this work, we discussed an algorithm that was flexible to general losses and could benefit from the Bernstein Condition. The biggest drawback of this algorithm is that it is not intractable. It might not be possible to propose a general and efficient implementation for IWAL-σ for any arbitrary hypothesis set. However, it lays the path for future work to propose efficient algorithms for a specific hypothesis class under a particular loss function, such as linear separators and squared loss.

Appendix A

Additional Information

A.1 Lemmas

Lemma A.1. *(Adapted from Lemma 3 of Kakade and Tewari (2009)) Suppose $X_1, X_2, \ldots, X_T$ is a martingale difference sequence with $|X_t| \leq b$. Let $V = \sigma^2 = \sum_{t=1}^{T} Var(X_t | X_1, \ldots, X_{t-1})$ be the sum of conditional variances of X_t's. Then we have, for any $\delta \leq 1/e$ and $T \geq 3$*

$$P(\sum_{t=1}^{T} X_t \geq \max\left(2\sigma\sqrt{\ln(\frac{4\ln(T)}{\delta})}, 3b\ln(\frac{4\ln(T)}{\delta})\right)) \leq \delta$$

The only difference between the adapted version of this Lemma here and the original Lemma in Kakade and Tewari (2009)'s work is that, in the original work on the right hand side of the inequality you will find $4\ln(T)\delta$ instead of δ alone. Here, we have only divided all the δs in the inequality by $4\ln(T)$ to achieve our desired form of the inequality. This inequality was first proposed by Freedman (1975).

Bibliography

Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228, 2013.

Maria-Florina Balcan and Ruth Urner. Active learning-modern learning theory., 2016.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72, 2006.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

Alina Beygelzimer, Daniel J Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, pages 199–207, 2010.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Region-based active learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2801–2809, 2019a.

Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, Ningshan Zhang, and Claudio Gentile. Active learning with disagreement graphs. In *International Conference on Machine Learning*, pages 1379–1387, 2019b.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Advances in neural information processing systems*, pages 337–344, 2005.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*, pages 235–242, 2006.

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

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

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518, 2019.

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

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

Steve Hanneke. Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3), 2014.

Steve Hanneke and Liu Yang. Negative results for active learning with convex losses. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 321–325, 2010.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 2755–2763, 2015.

Matti Kääriäinen. Generalization error bounds using unlabeled data. In *International Conference on Computational Learning Theory*, pages 127–142. Springer, 2005.

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

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

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Pascal Massart, Élodie Nédélec, et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein inequality for martingales: Application to online learning. 2013.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *International Conference on Machine Learning*, pages 3444–3452, 2017.

Tim van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Ningshan Zhang. Private communication. 2019.