UNIVERSITY OF CALIFORNIA SAN DIEGO

Algorithms for Query-Efficient Active Learning

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Songbai Yan

Committee in charge:

Professor Kamalika Chaudhuri, Co-Chair
Professor Tara Javidi, Co-Chair
Professor Sanjoy Dasgupta
Professor Farinaz Koushanfar
Professor Alon Orlitsky

2019

The dissertation of Songbai Yan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California San Diego

2019

DEDICATION

To my family.

EPIGRAPH

Nothing is more practical than a good theory.

—Vladimir N. Vapnik

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my two wonderful advisors Professor Kamalika Chaudhuri and Professor Tara Javidi. I am indebted to them for their tolerance and continuous support over the past five years, for sharing their insights and knowledge, teaching me how to conduct and present my research, and offering life and career suggestions. I could not have imagined having a better advisor and mentor for my PhD study. I would like to thank the rest of my doctoral committee: Professor Sanjoy Dasgupta, Professor Farinaz Koushanfar, and Professor Alon Orlitsky, for their insightful comments and encouragement.

I was very fortunate to intern at Microsoft Research and Amazon AWS, which provides me with valuable industrial research experience. I am very grateful to my mentors Chris Meek and Sheeraz Ahmed for their kind guidance and inspiring discussions.

I want to thank my fellow colleagues at University of California San Diego: Julaiti Alafate, Akshay Balsubramani, Joseph Geumlek, Shuang Liu, Suqi Liu, Stefanos Poulis, Shuang Song, Christopher Tosh, Sharad Vikram, Mengting Wan, Xinan Wang, Yizheng Wang, Jiapeng Zhang, and many others, for the stimulating discussions and for all the fun we have had. I would particularly like to thank Chicheng Zhang, who gently shared his knowledge in learning theory and gave me enormous constructive comments and suggestions for my research.

I am grateful to my undergraduate advisor Professor Liwei Wang for enlightening me the first glance of machine learning research and offering me career guidance. My sincere thanks also go to my friends met at Peking University, Kai Fan, Chi Jin, Hongyi Zhang, from whom I learned a lot in our discussion and reading group.

Last but not least, I would like to thank my parents and my wife Linjie Peng, for supporting me spiritually throughout writing this dissertation and my life in general. This dissertation would not have been possible without them.

Chapter 4 is based on the material in Advances in Neural Information Processing Systems 2017 (Songbai Yan and Chicheng Zhang, “Revisiting perceptron: Efficient and label-optimal active learning of halfspaces”). The dissertation author is the co-primary investigator and co-author of this material.

Chapter 5 is based on the materials in Allerton Conference on Communication, Control and Computing 2015 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning from Noisy and Abstention Feedback”) and Advances in Neural Information Processing Systems 2016 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning from Imperfect Labelers”). The dissertation author is the primary investigator and author of these materials.

Chapter 6 is based on the material in International Conference on Machine Learning 2018 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning with Logged Data”). The dissertation author is the co-primary investigator and co-author of this material.

Chapter 7 is based on the material submitted to Advances in Neural Information Processing Systems 2019 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”The Label Complexity of Active Learning from Observational Data”). The dissertation author is the co-primary investigator and co-author of this material.

VITA

2014	B. S. in Computer Science and Technology, Peking University, China
2017	M. S. in Computer Science, University of California San Diego, USA
2019	Ph. D. in Computer Science, University of California San Diego, USA

PUBLICATIONS

Songbai Yan, Kamalika Chaudhuri and Tara Javidi, “Active Learning with Logged Data”, *manuscript*, 2019.

Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha and Songbai Yan, “Exploring Connections Between Active Learning and Model Extraction”, *manuscript*, 2019.

Songbai Yan, Kamalika Chaudhuri and Tara Javidi, “Active Learning with Logged Data”, *International Conference on Machine Learning*, 2018.

Songbai Yan, Chicheng Zhang, “Revisiting Perceptron: Efficient and Label-Optimal Learning of Halfspaces”, *Neural Information Processing Systems*, 2017.

Songbai Yan, Kamalika Chaudhuri and Tara Javidi, “Active Learning from Imperfect Labelers”, *Neural Information Processing Systems*, 2016.

Songbai Yan, Kamalika Chaudhuri and Tara Javidi, “Active Learning from Noisy and Abstention Feedback”, *Allerton Conference on Communication, Control and Computing*, 2015.

ABSTRACT OF THE DISSERTATION

Algorithms for Query-Efficient Active Learning

by

Songbai Yan

Doctor of Philosophy in Computer Science

University of California San Diego, 2019

Professor Kamalika Chaudhuri, Co-Chair
Professor Tara Javidi, Co-Chair

Recent decades have witnessed great success of machine learning, especially for tasks where large annotated datasets are available for training models. However, in many applications, raw data, such as images, are abundant, but annotations, such as descriptions of images, are scarce. Annotating data requires human effort and can be expensive. Consequently, one of the central problems in machine learning is how to train an accurate model with as few human annotations as possible. Active learning addresses this problem by bringing the annotator to work together with the learner in the learning process. In active learning, a learner can sequentially select examples and ask the annotator for labels, so that it may require fewer annotations if the learning algorithm avoids querying less informative examples.

This dissertation focuses on designing provable query-efficient active learning algorithms. The main contributions are as follows. First, we study noise-tolerant active learning in the standard stream-based setting. We propose a computationally efficient algorithm for actively learning homogeneous halfspaces under bounded noise, and prove it achieves nearly optimal label complexity. Second, we theoretically investigate a novel interactive model where the annotator can not only return noisy labels, but also abstain from labeling. We propose an algorithm which utilizes abstention responses, and analyze its statistical consistency and query complexity under different conditions of the noise and abstention rate. Finally, we study how to utilize auxiliary datasets in active learning. We consider a scenario where the learner has access to a logged observational dataset where labeled examples are observed conditioned on a selection policy. We propose algorithms that effectively take advantage of both auxiliary datasets and active learning. We prove that these algorithms are statistically consistent, and achieve a lower label requirement than alternative methods theoretically and empirically.

Chapter 1

Introduction

Recent decades have witnessed great success of machine learning in various tasks, such as computer vision [RF17], natural language processing [VSP+17], and recommender systems [MTSVDH15]. One of the key factors to this success is the availability of large-scale annotated datasets such as images with class labels and products with user reviews. However, in many applications, raw data, such as DNA sequences and medical images, are abundant, but annotating them requires domain expertise and can be expensive. Consequently, one of the central problems in machine learning is how to train an accurate model with as few human annotations as possible.

One solution to this problem is through active learning where the learner works together with the annotator during the learning process. In active learning, a learner can sequentially select examples and ask the annotator for labels, so that it may require fewer annotations if the learning algorithm avoids querying less informative examples. It has been shown that active learning indeed helps reduce labeling efforts effectively in many tasks in natural language processing [SYL+18], computer vision [KSH+16, LWD+19], recommender systems [ERR16, KRG18], etc.

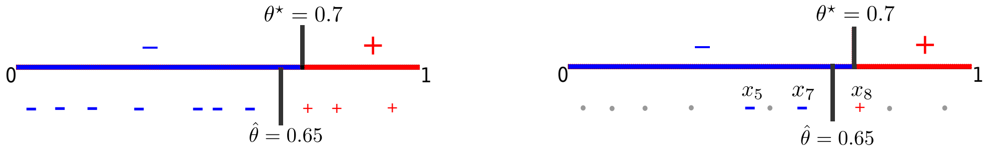


Figure 1.1: An example of learning a threshold $\theta^\star = 0.7$ with $n = 10$ unlabeled instances. Left: in supervised learning, all labels are requested, and a threshold $\hat\theta = 0.65$ is returned; Right: in active learning, the learner sequentially queries x_5, x_8, x_7, and returns $\hat\theta = 0.65$. Other instances (gray dots) are not queried.

How Active Learning Reduces Annotation Requirement

One classic example where active learning yields exponential label-efficiency improvement is learning a one-dimensional threshold with binary search. Consider a binary classification task where instances are real numbers from the unit interval $[0,1]$, and the label can be either positive $(+1)$ or negative (-1). Assume there is a threshold $\theta^\star \in [0,1]$ that perfectly separates the data, that is, any example x smaller than $\theta^\star$ is labeled negative and otherwise positive. To learn $\theta^\star$, in the standard supervised learning setting, as shown in Figure 1.1 (left), the learner would first draw n instances and request labels for all of them, and then output a threshold $\hat\theta$ that predicts correct labels for all observed instances. To guarantee $|\hat\theta - \theta^\star| \leq \varepsilon$, this passive learner needs $n = \Omega(\frac{1}{\varepsilon})$ labeled instances.

However, in the active learning setting, the learner could apply the binary search algorithm to find the threshold with much fewer labels. As shown in Figure 1.1 (right), the learner first draws n instances, but only requests the label for the instance in the middle. If its label is negative (resp. positive), then the learner can infer that all instances on its left (resp. right) are negative (resp. positive) as well and only needs to recursively search for $\theta^\star$ in the right (resp. left) half interval. In the end of this binary search procedure, the learner outputs a threshold $\hat\theta$ that predicts correct labels for all observed instances. It is easy to see that to guarantee $|\hat\theta - \theta^\star| \leq \varepsilon$, this active learner needs $n = O(\frac{1}{\varepsilon})$ unlabeled instances but only $O(\log\frac{1}{\varepsilon})$ labels, which is exponentially smaller than the label requirement for the passive learner.

The query strategy in this example, though seems simple, shares similar ideas with many general and widely used active learning algorithms, including generalized binary search [Now11] where instances that can rapidly narrow down the version space are selected are queried, margin-based methods [TK01, BBZ07] where examples near the current estimated boundary are queried, and disagreement based methods [CAL94, BBL06a] where examples are queried only if their labels cannot be confidently inferred.

How Active Learning Fails

Active learning is significantly more challenging in the nonrealizable case where no classifier in the hypothesis class achieves 100% accuracy. In this case, an improperly designed active learning algorithm may yield poor performance, mostly due to noisy annotations or sampling bias.

Noisy Annotations Human annotators can make mistakes, so the feedback returned to the learner may not always be consistent with the underlying ground truth. If handled improperly, an incorrect label may divert the active learning algorithm from the correct boundary and lead to a classifier with a high error rate. To illustrate this, consider again the threshold learning task in Figure 1.1. If the annotator returns an incorrect label +1 upon the first query x_5 from the active learner, and if the learner still uses the standard binary search algorithm, then the learner would incorrectly believe $\theta^\star \in [0, x_5]$, and recursively query in this half interval. Even if the annotator makes no more mistakes afterward, the learner will output a threshold $\hat{\theta} \leq 0.5$, which is far from the ground truth $\theta^\star$.

Sampling Bias An active learner often selects instances according to some criteria to query for labels. As a result, the distribution of labeled instances observed by the learner can be different from the actual data distribution, which can result in suboptimal solution especially in the non-realizable case where no classifier in the hypothesis class perfectly predicts all labels. To

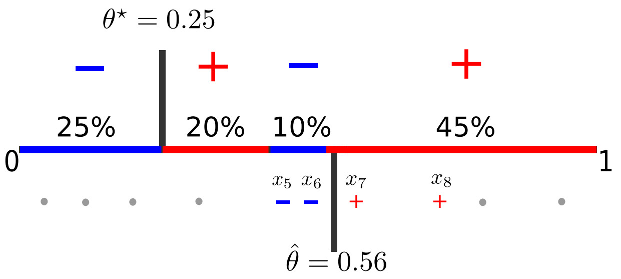


Figure 1.2: An example of learning thresholds in the non-realizable case. The best threshold is $\theta^\star = 0.25$ with error rate 0.1. In active learning, the learner sequentially queries x_5, x_8, x_7, x_6. Other instances (gray dots) are not queried, and a suboptimal threshold $\hat{\theta} = 0.56$ with error rate 0.21 is returned.

illustrate this, consider the threshold learning task as shown in Figure 1.2. We assume unlabeled instances are drawn uniformly from the unit interval and the corresponding ground truth labels are shown in the figure. In this example, no threshold function is 100% accurate, but we still want the algorithm to output a threshold as accurate as possible (in this example, the optimal threshold is $\theta^\star = 0.25$, which makes mistakes with probability 0.1). In supervised learning, the learner can apply the Empirical Risk Minimization principle: it first draws n instances, requests labels for all of them, and then outputs a threshold $\hat{\theta}$ that makes the fewest number of mistakes on observed instances. It can be shown that this method is *statistically consistent*, meaning that $\hat{\theta} \to \theta^\star$ as $n \to \infty$. However, the standard binary search algorithm for active learning is not statistically consistent: with high probability, it first queries an instance x_5 in the middle and receives a negative label; subsequently it recursively queries in its right and finally returns a classifier around 0.55, which is far from the optimal threshold $\theta^\star$ no matter how many labels are queried.

Hence, one main challenge in active learning is how to design query-efficient algorithms that tolerate mistakes of human annotators and guarantees statistical consistency. In the past decades, many active learning algorithms have been proposed and analyzed [CAL94, BBL06b, Han07, Das05, CN08, Now11, NJC15, CHK17, TD17, BBZ07, ABL14, ABHZ16, BDL09, BHLZ10, HAH$^+$15]. In this dissertation, we investigate three directions to advance the research on provably query-efficient active learning: (1) designing noise-tolerant active learning algorithms in the standard active learning setting; (2) exploring new interactive models beyond standard label feedback; (3) utilizing auxiliary information available to the learner.

Our Contributions

Efficient Active Learning of Halfspaces with Bounded Noise In Chapter 4, we study noise-tolerant learning of halfspaces under the standard stream-based active learning setting. We propose a computationally efficient Perceptron-based algorithm for actively learning homogeneous halfspaces under the uniform distribution over the unit sphere. We prove that under the bounded noise condition, where each label is flipped with probability at most $\frac{1}{2}$, our algorithm achieves a near-optimal label complexity.

Active Learning with Abstention Feedback In Chapter 5, we study a new interactive model where the annotator can not only return noisy labels, but also abstain from labeling. We consider different noise and abstention conditions of the annotator. We propose an algorithm which utilizes abstention responses. We prove this algorithm is statistically consistent and achieves nearly optimal query complexity under fairly natural conditions.

Active Learning with Logged Observational Data In the final two chapters, we study how to utilize an auxiliary dataset in active learning. In particular, We consider a scenario where the learner has access to a logged observational dataset where labeled examples are observed conditioned on a selection policy. In Chapter 6, we apply multiple importance sampling to utilize the logged data in active learning effectively and introduce a novel debiasing policy that selectively avoids querying those examples that are highly represented in the logged observational data. We prove that our algorithm is statistically consistent, and has a lower label requirement than alternatives both theoretically and empirically. In Chapter 7, we show how to apply variance control techniques to obtain a more sample-efficient error estimator, and then incorporate it into the active learning algorithm. We provably demonstrate that the new algorithm is statistically consistent as well as more label-efficient than the prior work.

Chapter 2

Related Work

2.1 Active Learning

In recent years, there has been extensive research in both theory and practice of active learning; see excellent surveys by [Set10, Das11, Han14]. On the theoretical side, many active learning algorithms have been proposed and analyzed. An incomplete list includes disagreement-based methods [CAL94, BBL06b, Han07], generalized binary search [Das05, CN08, Now11, NJC15, CHK17, TD17], margin-based methods [BBZ07, ABL14, ABHZ16], and importance weighted methods [BDL09, BHLZ10, HAH$^+$15]. There is also a considerable amount of work on lower bounds of label complexity for active learning under various noise conditions, and refined algorithms and analysis that approach these lower bounds [Han09, Kol10, RR11a, ZC14, HY15]. However, most of these algorithms are computationally efficient as they require either explicit enumeration of classifiers in hypothesis classes, or solving empirical 0-1 loss minimization problems.

On the practical side, many computationally efficient heuristics for active learning have been proposed, including uncertainty sampling [LG94, TK01], query by committee [SOS92, FSST97], maximizing expected model change [SCR08], and encouraging sample diversity [NS04, SS18]. It has been shown that these heuristics help reduce labeling efforts effectively in many tasks in natural language processing [SYL$^+$18], computer vision [KSH$^+$16, LWD$^+$19], recommender systems [ERR16, KRG18], etc.

In this dissertation, we advance the research on provably query-efficient active learning from three aspects: (1) we propose a label-optimal and computationally efficient active learning algorithm for learning halfspaces with bounded noise; (2) we explore a new interactive model that allows the annotator to abstain from labeling; (3) we show how to utilize an auxiliary observational dataset in active learning.

2.2 Efficient Learning of Halfspaces

Efficient learning of halfspaces is one of the central problems in machine learning [CST00]. In the realizable case, it is well known that linear programming finds a consistent hypothesis over data efficiently. In the nonrealizable setting, however, the problem is much more challenging.

A series of papers [ABSS93, FGKP06, GR09, KK14, Dan15] have shown the hardness of learning halfspaces with agnostic noise. The state of the art result [Dan15] shows that under standard complexity-theoretic assumptions, there exists a data distribution, such that the best linear classifier has error $o(1)$, but no polynomial time algorithms can achieve an error at most $\frac{1}{2} - \frac{1}{d^c}$ for every $c > 0$, even with improper learning. [KK14] shows that under standard assumptions (learning k-sparse parity with noise must have time $n^{\Omega(k)}$), even if the unlabeled distribution is Gaussian, any agnostic halfspace learning algorithm must run in time $(\frac{1}{\varepsilon})^{\Omega(\ln d)}$ to achieve an excess error of ε. These results indicate that, to have nontrivial guarantees on learning halfspaces with noise in polynomial time, one has to make additional assumptions on the data distribution over instances and labels.

Since it is believed to be hard for learning halfspaces in the general agnostic setting, it is natural to consider algorithms that work under more moderate noise conditions. Despite considerable efforts, there are only a few halfspace learning algorithms that are both computationally-efficient and label-efficient. In the realizable setting, [DKM05, BBZ07, BL13] propose computationally efficient active learning algorithms which have an optimal label complexity of $\tilde{O}(d \ln \frac{1}{\varepsilon})$. Under the bounded noise setting [MN06], the only known algorithms that are both label-efficient and computationally-efficient are [ABHU15, ABHZ16]. [ABHU15] uses a margin-based framework which queries the labels of examples near the decision boundary. To achieve computational efficiency, it adaptively chooses a sequence of hinge loss minimization problems to optimize as opposed to directly optimizing the 0-1 loss. It works only when the label flipping probability upper bound η is small ($\eta \leq 1.8 \times 10^{-6}$). [ABHZ16] improves over [ABHU15] by adapting a polynomial regression procedure into the margin-based framework. It works for any $\eta < 1/2$, but its label complexity is $O(d^{O(\frac{1}{(1-2\eta)^4})} \ln \frac{1}{\varepsilon})$, which is far worse than the information-theoretic lower bound $\Omega(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\varepsilon})$. Recently [CHK17] gives an efficient algorithm with a near-optimal label complexity under the membership query model where the learner can query on synthesized points. However, it is unclear how to transform the DC algorithm in [CHK17] into a computationally efficient stream-based active learning algorithm where the learner can only query on points drawn from the data distribution.

In Chapter 4, we provide a Perceptron-based algorithm that is computationally efficient and achieves nearly optimal label complexity for learning halfspaces under the bounded noise setting.

2.3 Interactive Models for Active Learning

In the standard active learning setting, the learner obtains labels from an annotator. Three interactive models between the learner and the annotator are commonly used: (1) the membership query model, where the learner can query any instances in the instance space for labels; (2) the stream-based query model, where the learner is presented a stream of unlabeled instances drawn from an underlying distribution one at a time, and for each of them the learner needs to decide whether to query for its label or not in an online fashion; (3) the pool-based query model, where the learner is presented a pool of unlabeled examples drawn from an underlying distribution, and it can iteratively query some of them for labels. In Chapter 5, we work with the membership query model. In Chapters 4, 6, and 7, we work with the stream-based query model. We note that an algorithm for the stream-based query model also works under the pool-based query model, while converting an algorithm from the pool-based model to stream-based model is nontrivial and there can be a significant gap with respect to label complexity under these two models [SH16].

Many novel interactive models are studied where annotators can provide information beyond label feedback. For example, [Ang88, Heg95] consider equivalence query where the learner presents a classifier to the annotator, and the annotator either confirms this classifier is correct or otherwise returns a counter-example. [BH12] considers class-conditional query where the learner presents an unlabeled instance set U and a class label, and the annotator returns an example of class c from U. [ZC15] considers a setting where the learner can choose to query for labels from a cheap but noisy annotator or an expensive but accurate one. [BHLZ16] considers search query where the learner presents a set of classifiers V, and the annotator returns a labeled example on which all classifiers in V predict incorrectly. [XZM^{+}17] considers pairwise comparison query where the learner presents two unlabeled examples, and the annotator returns which one is more likely to be positive.

In Chapter 5, in addition to providing possibly noisy labels, we allow the annotator to abstain from labeling. [FZ12, KFR$^+$15] consider learning with abstention feedback in computer vision applications, but they only propose heuristic query strategies and do not provide any theoretical guarantees. In our work, we rigorously show when abstention feedback helps active learning, and provide an algorithm that achieves the nearly optimal query complexity.

2.4 Learning with Observational Data

Learning from logged observational data is a fundamental problem in machine learning with applications to causal inference [SJS17], information retrieval [SLLK10, LCKG15, HLR16], recommender systems [LCLS10, SSS$^+$16], online learning [AHK$^+$14, WAD17], and reinforcement learning [Tho15, TTG15, MLBP16]. This problem is also closely related to covariate shift [Zad04, SKM07, BDBC$^+$10] in domain adaptation.

When the logging policy is *unknown*, the direct method [DLL11] finds a classifier using observed data. This method, however, is vulnerable to the sample selection bias [HLR16, JSS16]. Existing de-biasing procedures include tree-based methods to partition the data space [AI16, Kal17], and learning good representations with deep neural networks to align the observational and population data [JSS16, SJS17].

When the logging policy is *known*, we can learn a classifier by optimizing a loss function that is an unbiased estimator of the expected error rate. The most common estimator is the importance weighted estimator that reweights examples according to inverse propensity scores [RR83]. This method is unbiased when propensity scores are accurate, but may have a high variance when some propensity scores are close to zero. To resolve this, [BPQC$^+$13, SLLK10, SJ15a] propose to truncate the inverse propensity score, [SJ15b] proposes to use normalized importance sampling, [MP09, SJ15a] propose to add a regularizer based on empirical variance to the loss function to favor models with low loss variance, [JL16, DLL11, TB16, WAD17] propose doubly robust estimators, and recently [TTG15, ABSJ17] suggest adjusting importance weights according to data to reduce the variance further.

Most existing work on learning with observational data falls into the passive learning paradigm, that is, they first collect the observational data and then train a classifier. To the best of our knowledge, there is no prior work with theoretical guarantees that combines passive and active learning with a logged observational dataset. [BDL09] considers active learning with warm-start where the algorithm is presented with a labeled dataset prior to active learning, but the labeled dataset is not observational: it is assumed to be drawn from the same distribution for the entire population. [AZvdS19] and [SSS+19] consider active learning for predicting individual treatment effects, which is similar to our task. They take a Bayesian approach which does not need to know the logging policy, but assumes the true model is from a known distribution family. Additionally, they do not provide label complexity bounds. A related line of research considers active learning for domain adaptation, and they are mostly based on heuristics [SRD+11, ZJL+16], utilizing a clustering structure [KGR+15], or non-parametric methods [KM18]. In other related settings, [ZAI+19] considers warm-starting contextual bandits targeting at minimizing the cumulative regret instead of the final prediction error; [KAH+17] studies active learning with bandit feedback without any logged observational data.

In Chapter 6, we provide an active learning algorithm that utilizes the logged observational data to reduce the number of label queries with theoretical guarantees. In Chapter 7, we improve this algorithm by incorporating a more efficient variance-controlled importance sampling into active learning and show that it leads to a better label complexity.

Chapter 3

Preliminaries

3.1 Learning Scenarios

This dissertation focuses on binary classification tasks in machine learning. In this task, we assume examples to be classified come from an instance space $\mathcal{X}$, and the classification outcome belongs to a binary label space $\mathcal{Y}$. The output of a learning algorithm is a classifier (also known as a hypothesis), which is a function $h : \mathcal{X} \to \mathcal{Y}$ that given an instance predicts its label. We restrict the output of the learning algorithm to classifiers from a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$.

We consider two active learning scenarios: stream-based active learning, and active learning with membership queries. In the following two subsections, we explain how instances and labels are generated, how the algorithm interacts with the annotator, and how the performance of the learning algorithm is evaluated in each scenario.

3.1.1 Stream-Based Active Learning

Stream-based active learning uses the Probably Approximately Correct (PAC) learning framework [Val84]. In this setting, there is an underlying distribution D over $\mathcal{X} \times \mathcal{Y}$. At time $t = 1, 2, \ldots$, an independent and identically distributed (i.i.d.) example (X_t, Y_t) is drawn from D, and only X_t is presented to the learner. The learner can decide whether to query for Y_t, and it observes Y_t only if it chooses to query. This decision can depend on all instances up to time t and previously observed labels.

In this setting, the performance of a classifier h is measured by the 0-1 loss $l(h) := \mathbb{P}_D(h(X) \neq Y)$. The performance of a learning algorithm is measured by *query complexity* which is the number of queries needed to guarantee a certain loss. In particular, for any algorithm $\mathcal{A}$, excess error ε, and confidence level δ, the query complexity $\Lambda(\mathcal{A}, \varepsilon, \delta)$ is defined as the minimum number of label queries such that $\mathcal{A}$ outputs a classifier h satisfying $l(h) \leq \min_{h' \in \mathcal{H}} l(h') + \varepsilon$ with probability at least $1 - \delta$ after querying this number of labels.

3.1.2 Active Learning with Membership Queries

In active learning with membership queries, the learner can synthesize an instance in $\mathcal{X}$ to query for the label. At time $t = 1, 2, \ldots$, the learner chooses an instance $X_t \in \mathcal{X}$ and queries the labeler. The response of the labeler follows an underlying conditional distribution $D_{Y|X}$. For each queried instance X_t, the labeler draws an i.i.d. label Y_t from $D_{Y|X=X_t}$ and returns it to the learner.

In this setting, we assume there is an underlying optimal classifier $h^\star \in \mathcal{H}$ and a metric $d : \mathcal{H} \times \mathcal{H} \to [0, \infty)$. The performance of a classifier h is measured by its distance to the optimal classifier $d(h^\star, h)$. Similar to the stream-based setting, the performance of a learning algorithm is also measured by *query complexity*. In the membership query setting, for any algorithm $\mathcal{A}$, error

ε, and confidence level δ, the query complexity $\Lambda(\mathcal{A},\varepsilon,\delta)$ is defined as the minimum number of label queries such that $\mathcal{A}$ outputs a classifier h satisfying $d(h^\star,h) \le \varepsilon$ with probability at least $1-\delta$ after querying this number of labels.

3.2 Definitions

Let $\mathbb{1}[A]$ be the indicator function: $\mathbb{1}[A] = 1$ if A is true, and 0 otherwise. For $x = (x_1,\dots,x_d) \in \mathbb{R}^d$ ($d > 1$), denote $(x_1,\dots,x_{d-1})$ by $\tilde{x}$. Define $\ln x := \log_e x$, and $[\ln\ln]_+(x) = \ln\ln\max\{x, e^e\}$. Define $\tilde{O}(f(\cdot)) = O(f(\cdot)\log f(\cdot))$, and $\tilde{\Omega}(f(\cdot)) = \Omega(f(\cdot)/\ln f(\cdot))$. We say $g(\cdot) = \tilde{\Theta}(f(\cdot))$ if and only if $g(\cdot) = \tilde{O}(f(\cdot))$ and $g(\cdot) = \tilde{\Omega}(f(\cdot))$

Definition 3.1. Suppose $\gamma \ge 1$. A function $g : [0,1]^d \to \mathbb{R}$ is (K,γ)-*Hölder smooth*, if it is continuously differentiable up to $\lfloor\gamma\rfloor$-th order, and for any $x,y \in [0,1]^d$, $\left|g(y) - \sum_{m=0}^{\lfloor\gamma\rfloor} \frac{\partial^m g(x)}{m!}(y-x)^m\right| \le K\|y-x\|^\gamma$. We denote this class of functions by $\Sigma(K,\gamma)$.

Definition 3.2. For any conditional distribution $D_{Y|X}$, the *Bayes Optimal Classifier* h_{Bayes} is defined as $h_{\text{Bayes}}(x) = +1$ if $\mathbb{P}_D(Y = +1 \mid X = x) > \frac{1}{2}$ else -1.

Next, we introduce some standard definitions in the PAC framework for stream-based active learning. Unless otherwise specified, all probabilities and expectations are over the distribution D.

Define the optimal classifier $h^\star := \arg\min_{h\in\mathcal{H}} l(h)$, and the optimal error $\nu := l(h^\star)$. If $\nu = 0$,we are said to be in the *realizable* case as there is a classifier $h^\star$ in $\mathcal{H}$ that predicts all labels correctly. If we make no assumption on the data distribution D, we are said to be in the *agnostic* case.

Recall the expected error rate $l(h) = \mathbb{P}(h(X) \neq Y)$. For $S = \{(X_1,Y_1),\dots,(X_n,Y_n)\} \subset$

$\mathcal{X} \times \mathcal{Y}$, define the empirical error $l(h,S) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[h(X_i) \neq Y_i]$. Additionally, define $\rho(h_1,h_2) := \mathbb{P}(h_1(X) \neq h_2(X))$ to be the disagreement probability mass between h_1 and h_2, and $\rho_S(h_1,h_2) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[h_1(X_i) \neq h_2(X_i))]$ for $S = \{X_1,X_2,\ldots,X_n\} \subset \mathcal{X}$ to be the empirical disagreement mass between h_1 and h_2 on S.

For any $h \in \mathcal{H}$, $r > 0$, define $B(h,r) := \{h' \in \mathcal{H} \mid \rho(h,h') \leq r\}$ to be r-ball around h. For any $C \subseteq \mathcal{H}$, define the disagreement region $\mathrm{DIS}(C) := \{x \in \mathcal{X} \mid \exists h_1 \neq h_2 \in C \text{ s.t. } h_1(x) \neq h_2(x)\}$.

Definition 3.3. For any $r > 0$, define $\theta(r) := \sup_{r' > r} \frac{1}{r'}\mathbb{P}(\mathrm{DIS}(B(h^\star, r')))$ to be the *disagreement coefficient*. Define $\theta := \theta(2\nu)$.

Finally, we introduce some definitions on distributions.

Definition 3.4. Let $\mathbb{P},\mathbb{Q}$ be two probability measures on a common measurable space and $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$.

- The KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as $D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{X\sim\mathbb{P}} \ln \frac{\mathbb{P}(X)}{\mathbb{Q}(X)}$.
- We define $d_{\mathrm{KL}}(p,q) = D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q})$, where $\mathbb{P},\mathbb{Q}$ are distributions of a Bernoulli(p) and a Bernoulli(q) random variables respectively.
- For random variables X,Y,Z, define the mutual information between X and Y under $\mathbb{P}$ as $I(X;Y) = D_{\mathrm{KL}}\left(\mathbb{P}(X,Y), \mathbb{P}(X)\mathbb{P}(Y)\right) = \mathbb{E}_{X,Y} \ln \frac{\mathbb{P}(X,Y)}{\mathbb{P}(X)P(Y)}$, and define the mutual information between X and Y conditioned on Z under $\mathbb{P}$ as $I(X;Y \mid Z) = \mathbb{E}_{X,Y,Z} \ln \frac{\mathbb{P}(X,Y|Z)}{\mathbb{P}(X|Z)P(Y|Z)}$.
- For a random variable sequence $X_1,X_2,\ldots$, denote by X^n the subsequence $\{X_1,X_2,\ldots X_n\}$.

3.3 The Disagreement-Based Active Learning Algorithm

The Disagreement-Based Active Learning (DBAL) algorithm, shown as Algorithm 1, is a general active learning algorithm that has rigorous theoretical guarantees and can be implemented practically. It is first proposed by [CAL94] in the realizable case and then improved by [BBL06b] to work in the general agnostic case. A survey can be found in [Han14].

Algorithm 1 Standard Disagreement-Based Active Learning Algorithm

```
Input: confidence $\delta$, number of unlabeled examples $n$
Request a labeled example $(X_1, Y_1)$
$\tilde{S} \leftarrow \{(X_1, Y_1)\}$; $C_0 \leftarrow \mathcal{H}$;$K \leftarrow \log_2 n$
for $k = 1, \ldots, K$ do
    $\hat{h}_{k-1} \leftarrow \arg\min_{h \in C_{k-1}} l(h, \tilde{S})$, $\delta_k \leftarrow \frac{\delta}{k(k+1)}$
    for $t = 2^k$ to $2^{k+1} - 1$ do
        Draw an unlabeled instance $X_t$
        if $X_t \in \mathrm{DIS}(C_{k-1})$ then
            Query for its label $\tilde{Y}_t \leftarrow Y_t$
        else
            Infer its label $\tilde{Y}_t \leftarrow \hat{h}_{k-1}(X_t)$.
        end if
        $\tilde{S} \leftarrow \tilde{S} \cup \{(X_t, \tilde{Y}_t)\}$
    end for
    Update the candidate set $C_k \leftarrow \{h \in C_{k-1} \mid l(h, \tilde{S}) \leq l(\hat{h}_{k-1}, \tilde{S}) + U(h, \hat{h}_{k-1}, \tilde{S}, \delta_k)\}$
end for
Output $\hat{h} = \arg\min_{h \in C_K} l(h, \tilde{S})$
```

DBAL iteratively maintains a candidate set of classifiers C_k to be the confidence set of the optimal classifier $h^\star$. At the k-th iteration, the learner draws 2^k unlabeled examples. For each instance X_t among them, if it falls into the current disagreement region $\mathrm{DIS}(C_{k-1})$, meaning that there are at least two classifiers in C_{k-1} that predict different labels on X_t, then the algorithm queries for its label Y_t; otherwise, it infers the label as $\tilde{Y} = \hat{h}_{k-1}(X)$. In the end of each iteration, the queried and inferred labels are used to shrink the candidate set.

It has been shown that Algorithm 1 with a proper choice of $U(\cdot)$ achieves a label complexity of $\tilde{O}\left(\theta\frac{\nu^2}{\varepsilon^2}d\log\frac{1}{\delta}\right)$ where d is the VC dimension [Vap98] of $\mathcal{H}$, which is always no worse than the minimiax label complexity $\tilde{\Theta}(\frac{\nu+\varepsilon}{\varepsilon^2}(d+\log\frac{1}{\delta}))$ for passive learning [Han14].

Chapter 4

Efficient Active Learning of Halfspaces with Bounded Noise

4.1 Introduction

In this chapter, we study the problem of designing efficient noise-tolerant algorithms for actively learning homogeneous halfspaces in the streaming setting. We are given access to a data distribution from which we can draw unlabeled examples, and a noisy labeler O that we can query for labels. The goal is to find a computationally efficient algorithm to learn a halfspace that best classifies the data while making as few queries to the labeler as possible.

There has been a large body of work on the theory of active learning, showing sharp distribution-dependent label complexity bounds [CAL94, BBL09, Han07, DHM07, Han09, Kol10, ZC14, HAH$^+$15]. However, most of these general active learning algorithms rely on solving empirical risk minimization problems, which are computationally hard in the presence of noise [ABSS93].

On the other hand, existing computationally efficient algorithms for learning halfspaces [BFKV98, DV04, KKMS08, KLS09, ABL14, Dan15, ABHU15, ABHZ16] are not optimal in terms of label requirements. These algorithms have different degrees of noise tolerance (e.g. adversarial noise [ABL14], malicious noise [KL93], random classification noise [AL88], bounded noise [MN06], etc), and run in time polynomial in $\frac{1}{\varepsilon}$ and d. Some of them naturally exploit the utility of active learning [ABL14, ABHU15, ABHZ16], but they do not achieve the sharpest label complexity bounds in contrast to those computationally-inefficient active learning algorithms [BBZ07, BL13, ZC14].

Therefore, a natural question is: is there any active learning halfspace algorithm that is computationally efficient, and has a minimum label requirement? This has been posed as an open problem in [Mon06]. In the realizable setting, [DKM05, BBZ07, BL13, TD17] give efficient algorithms that have optimal label complexity of $\tilde{O}(d \ln \frac{1}{\varepsilon})$ under some distributional assumptions. However, the challenge still remains open in the nonrealizable setting. It has been shown that learning halfspaces with agnostic noise even under Gaussian unlabeled distribution is hard [KK14]. Nonetheless, under the bounded noise condition, we propose a Perceptron-based algorithm which is computationally efficient, and achieves near-optimal label complexity bound. In addition, this algorithm can be converted to a passive learning algorithm that has near optimal sample complexities.

4.2 Setup

We consider learning homogeneous halfspaces under uniform distribution over the unit sphere. The instance space $\mathcal{X}$ is the unit sphere in $\mathbb{R}^d$, which we denote by $\mathbb{S}^{d-1} := \left\{ x \in \mathbb{R}^d : \|x\| = 1 \right\}$. We assume $d \geq 3$ throughout this chapter. The label space $\mathcal{Y} = \{+1, -1\}$. We assume all data points (x, y) are drawn i.i.d. from an underlying distribution D over $\mathcal{X} \times \mathcal{Y}$.

We denote by $D_{\mathcal{X}}$ the marginal of D over $\mathcal{X}$ (which is uniform over $\mathbb{S}^{d-1}$), and $D_{Y|X}$ the conditional distribution of Y given X. Our algorithm is allowed to draw unlabeled examples $x \in \mathcal{X}$ from $D_{\mathcal{X}}$, and to make queries to a labeler O for labels. Upon query x, O returns a label y drawn from $D_{Y|X=x}$. The hypothesis class of interest is the set of homogeneous halfspaces $\mathcal{H} := \left\{ h_w(x) = \operatorname{sign}(w \cdot x) \mid w \in \mathbb{S}^{d-1} \right\}$. For any hypothesis $h \in \mathcal{H}$, we define its error rate $l(h) := \mathbb{P}_D[h(X) \neq Y]$. We will drop the subscript D in $\mathbb{P}_D$ when it is clear from the context. Given a dataset $S = \left\{(X_1, Y_1), \ldots, (X_m, Y_m)\right\}$, we define the empirical error rate of h over S as $l_S(h) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left\{h(x_i) \neq y_i\right\}$.

Definition 4.1 (Bounded Noise [MN06])**.** We say that the labeler O satisfies the η-*bounded noise condition* for some $\eta \in [0, 1/2)$ with respect to u, if for any x, $\mathbb{P}[Y \neq \operatorname{sign}(u \cdot x) \mid X = x] \leq \eta$.

It can be seen that under η-bounded noise condition, h_u is the Bayes classifier.

For two unit vectors v_1, v_2, denote by $\theta(v_1, v_2) = \arccos(v_1 \cdot v_2)$ the angle between them. The following lemma gives relationships between errors and angles (see also Lemma 1 in [ABHZ16]).

Lemma 4.2. *For any $v_1, v_2 \in \mathbb{S}^{d-1}$, $\left|l(h_{v_1}) - l(h_{v_2})\right| \leq \mathbb{P}\left[h_{v_1}(X) \neq h_{v_2}(X)\right] = \frac{\theta(v_1, v_2)}{\pi}$.*

Additionally, if the labeler satisfies the η-bounded noise condition with respect to u, then for any vector v, $\left|l(h_v) - l(h_u)\right| \geq (1 - 2\eta)\mathbb{P}\left[h_v(X) \neq h_u(X)\right] = \frac{1-2\eta}{\pi}\theta(v, u)$.

Given access to unlabeled examples drawn from $D_{\mathcal{X}}$ and a labeler O, our goal is to find a polynomial time algorithm $\mathcal{A}$ such that with probability at least $1 - \delta$, $\mathcal{A}$ outputs a halfspace $h_v \in \mathcal{H}$ with $\mathbb{P}[\operatorname{sign}(v \cdot X) \neq \operatorname{sign}(u \cdot X)] \leq \varepsilon$ for some target accuracy ε and confidence δ. (By Lemma 4.2, this guarantees that the excess error of h_v is at most ε, namely, $l(h_v) - l(h_u) \leq \varepsilon$.) The desired algorithm should make as few queries to the labeler O as possible.

We say an algorithm $\mathcal{A}$ achieves a *label complexity* of $\Lambda(\varepsilon, \delta)$, if for any target halfspace

$h_u \in \mathcal{H}$, with probability at least $1-\delta$, $\mathcal{A}$ outputs a halfspace $h_v \in \mathcal{H}$ such that $l(h_v) \leq l(h_u)+\varepsilon$, and requests at most $\Lambda(\varepsilon,\delta)$ labels from labeler O.

4.3 Algorithm

Our main algorithm, Algorithm 2, works in epochs. It works under the bounded noise model, if its sample schedule $\{m_k\}$ and band width $\{b_k\}$ are set appropriately with respect to each noise model. At the beginning of each epoch k, it assumes an upper bound of $\frac{\pi}{2^k}$ on $\theta(v_{k-1},u)$, the angle between current iterate v_{k-1} and the underlying halfspace u. As we will see, this can be shown to hold with high probability inductively. Then, it calls procedure MODIFIED-PERCEPTRON (Algorithm 3) to find an new iterate v_k, which can be shown to have an angle with u at most $\frac{\pi}{2^{k+1}}$ with high probability. The algorithm ends when a total of $k_0 = \lceil \log_2 \frac{1}{\varepsilon} \rceil$ epochs have passed.

For simplicity, we assume for the rest of the chapter that the angle between the initial halfspace v_0 and the underlying halfspace u is acute, that is, $\theta(v_0,u) \leq \frac{\pi}{2}$; Appendix A.2 shows that this assumption can be removed with a constant overhead in terms of label and time complexities.

Algorithm 2 ACTIVE-PERCEPTRON

```
Input: Labeler $O$, initial halfspace $v_0$, target error $\varepsilon$, confidence $\delta$, sample schedule $\{m_k\}$, band width $\{b_k\}$.
Output: learned halfspace $v$.
Let $k_0 = \lceil \log_2 \frac{1}{\varepsilon} \rceil$.
for $k = 1,2,\ldots,k_0$ do
    $v_k \leftarrow$ MODIFIED-PERCEPTRON$(O, v_{k-1}, \frac{\pi}{2^k}, \frac{\delta}{k(k+1)}, m_k, b_k)$.
end for
Return $v_{k_0}$.
```

Procedure MODIFIED-PERCEPTRON (Algorithm 3) is the core component of Algorithm 2. It sequentially performs a modified Perceptron update rule on the selected new examples (x_t, y_t) [MS54, BFKV98, DKM05]:

$$w_{t+1} \leftarrow w_t - 2\mathbb{1}\left\{y_t w_t \cdot x_t < 0\right\}(w_t \cdot x_t) \cdot x_t \tag{4.1}$$

Define $\theta_t := \theta(w_t, u)$. Update rule (4.1) implies the following relationship between θ_{t+1} and θ_t (See Lemma 4.12 for its proof):

$$\cos\theta_{t+1} - \cos\theta_t = -2\mathbb{1}\left\{y_t w_t \cdot x_t < 0\right\}(w_t \cdot x_t) \cdot (u \cdot x_t) \tag{4.2}$$

This motivates us to take $\cos\theta_t$ as our measure of progress; we would like to drive $\cos\theta_t$ up to 1(so that θ_t goes down to 0) as fast as possible.

To this end, MODIFIED-PERCEPTRON samples new points x_t under time-varying distributions $D_X|_{R_t}$ and query for their labels, where $R_t = \left\{x \in \mathbb{S}^{d-1} : \frac{b}{2} \leq w_t \cdot x \leq b\right\}$ is a band inside the unit sphere. The rationale behind the choice of R_t is twofold:

1. We set R_t to have a probability mass of $\tilde{\Omega}(\varepsilon)$, so that the time complexity of rejection sampling is at most $\tilde{O}(\frac{1}{\varepsilon})$ per example. Moreover, in the adversarial noise setting, we set R_t large enough to dominate the noise of magnitude $\nu = \tilde{\Omega}(\varepsilon)$.

2. Unlike the active Perceptron algorithm in [DKM05] or other margin-based approaches (for example [TK01, BBZ07]) where examples with small margin are queried, we query the label of the examples with a range of margin $[\frac{b}{2}, b]$. From a technical perspective, this ensures that θ_t decreases by a decent amount in expectation (see Lemma 4.13 for details).

Following the insight of [GCB09], we remark that the modified Perceptron update (4.1) on distribution $D_X|_{R_t}$ can be alternatively viewed as performing stochastic gradient descent on a special non-convex loss function $\ell(w,(x,y)) = \min(1, \max(0, -1 - \frac{2}{b} y w \cdot x))$. It is an interesting open question whether optimizing this new loss function can lead to improved empirical results for learning halfspaces.

Algorithm 3 MODIFIED-PERCEPTRON

```
Input: Labeler $O$, initial halfspace $w_0$, angle upper bound $\theta$, confidence $\delta$, number of iterations $m$, band width $b$.
Output: Improved halfspace $w_m$.
for $t = 0, 1, 2, \ldots, m-1$ do
    Define region $R_t = \left\{x \in \mathbb{S}^{d-1} : \frac{b}{2} \le w_t \cdot x \le b\right\}$.
    Rejection sample $x_t \sim D_X|_{R_t}$. In other words, draw $x_t$ from $D_X$ until $x_t$ is in $R_t$. Query $O$ for its label $y_t$.
    $w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t w_t \cdot x_t < 0\} \cdot (w_t \cdot x_t) \cdot x_t$.
end for
Return $w_m$.
```

4.4 Analysis

We show that Algorithm 2 works in the bounded noise model, achieving computational efficiency and near-optimal label complexity. To this end, we first establish a lower bound on the label complexity under bounded noise, and then give computational and label complexity upper bounds.

4.4.1 Lower Bounds

We first present an information-theoretic lower bound on the label complexity in the bounded noise setting under uniform distribution. This extends the distribution-free lower bounds of [RR11a, Han14], and generalizes the realizable-case lower bound of [KMT93] to the bounded noise setting. Our lower bound can also be viewed as an extension of [WS16]'s Theorem 3; specifically it addresses the hardness under the α-Tsybakov noise condition where $\alpha = 0$ (while [WS16]'s Theorem 3 provides lower bounds when $\alpha \in (0,1)$).

Theorem 4.3. *For any $d > 4$, $0 \leq \eta < \frac{1}{2}$, $0 < \varepsilon \leq \frac{1}{4\pi}$, $0 < \delta \leq \frac{1}{4}$, for any active learning algorithm $\mathcal{A}$, there is a $u \in \mathbb{S}^{d-1}$, and a labeler O that satisfies η-bounded noise condition with respect to u, such that if with probability at least $1-\delta$, $\mathcal{A}$ makes at most n queries of labels to O and outputs $v \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\operatorname{sign}(v \cdot X) \neq \operatorname{sign}(u \cdot X)] \leq \varepsilon$, then $n \geq \Omega\left(\frac{d \log \frac{1}{\varepsilon}}{(1-2\eta)^2} + \frac{\eta \log \frac{1}{\delta}}{(1-2\eta)^2}\right)$.*

Theorem 4.3 is proved with techniques from information theory. We will use the following two folklore information-theoretic lower bounds.

Lemma 4.4. *Let $\mathcal{W}$ be a class of parameters, and $\{P_w : w \in \mathcal{W}\}$ be a class of probability distributions indexed by $\mathcal{W}$ over some sample space $\mathcal{X}$. Let $d : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a semi-metric. Let $\mathcal{V} = \{w_1, \ldots, w_M\} \subseteq \mathcal{W}$ such that $\forall i \neq j$, $d(w_i, w_j) \geq 2s > 0$. Let V be a random variable uniformly taking values from $\mathcal{V}$, and X be drawn from P_V. Then for any algorithm $\mathcal{A}$ that given a sample X drawn from P_w outputs $\mathcal{A}(X) \in \mathcal{W}$, the following inequality holds:*

$$\sup_{w \in \mathcal{W}} P_w\left(d(w, \mathcal{A}(X)) \geq s\right) \geq 1 - \frac{I(V;X) + \ln 2}{\ln M}$$

Proof. For any algorithm $\mathcal{A}$, define a test function $\hat{\Psi} : \mathcal{X} \to \{1, \ldots, M\}$ such that

$$\hat{\Psi}(X) = \arg \min_{i \in \{1,\ldots,M\}} d(\mathcal{A}(X), w_i)$$

We have

$$\sup_{w \in \mathcal{W}} P_w\left(d(w, \mathcal{A}(X)) \geq s\right) \geq \max_{w \in \mathcal{V}} P_w\left(d(w, \mathcal{A}(X)) \geq s\right) \geq \max_{i \in \{1,\ldots,M\}} P_{w_i}\left(\hat{\Psi}(X) \neq i\right)$$

The desired result follows by classical Fano's Inequality:

$$\max_{i\in\{1,\ldots,M\}} P_{w_i}\left(\hat{\Psi}(X)\neq i\right)\geq 1-\frac{I(V;X)+\ln 2}{\ln M}$$

□

Lemma 4.5. *[AB09, Lemma 5.1] Let* $\gamma\in(0,1)$*,* $\delta\in(0,\frac{1}{4})$*,* $p_0=\frac{1-\gamma}{2}$*,* $p_1=\frac{1+\gamma}{2}$*. Suppose that* $\alpha\sim$*Bernoulli*$(\frac{1}{2})$ *is a random variable,* $\xi_1,\ldots,\xi_m$ *are i.i.d. (given* α*) Bernoulli*(p_α) *random variables. If* $m\leq 2\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\frac{1}{8\delta(1-2\delta)}\right\rfloor$*, then for any function* $f:\{0,1\}^m\to\{0,1\}$*, we have* $\mathbb{P}\left(f(\xi_1,\ldots,\xi_m)\neq\alpha\right)>\delta$*.*

Next, we present two technical lemmas.

Lemma 4.6. *[Lon95, Lemma 6] For any* $0<\gamma\leq\frac{1}{2}$*,* $d\geq 1$*, there is a finite set* $\mathcal{V}\in\mathbb{S}^{d-1}$ *such that the following two statements hold:*

1. *For any distinct* $w_1,w_2\in\mathcal{V}$*,* $\theta(w_1,w_2)\geq\pi\gamma$*;*

2. $|\mathcal{V}|\geq\frac{\sqrt{d}}{2}\left(\frac{1}{2\pi\gamma}\right)^{d-1}-1$.

Lemma 4.7. *If* $p\in[0,1]$ *and* $q\in(0,1)$*, then* $d_{KL}(p,q)\leq\frac{(p-q)^2}{q(1-q)}$*.*

Proof.

$$\begin{aligned}
d_{\mathrm{KL}}(p,q) &= p\ln\frac{p}{q}+(1-p)\ln\frac{1-p}{1-q}\\
&\leq p(\frac{p}{q}-1)+(1-p)(\frac{1-p}{1-q}-1)\\
&= \frac{(p-q)^2}{q(1-q)}
\end{aligned}$$

where the inequality follows by $\ln x\leq x-1$. □

Now, Theorem 4.3 is immediate from the following two lemmas.

Lemma 4.8. *For any* $0 \leq \eta < \frac{1}{2}$, $d > 4$, $0 < \varepsilon \leq \frac{1}{4\pi}$, $0 < \delta < \frac{1}{2}$, *for any active learning algorithm* $\mathcal{A}$, *there is a* $u \in \mathbb{S}^{d-1}$, *and a labeler* O *that satisfies* η*-bounded noise condition with respect to* u, *such that if with probability at least* $1-\delta$, $\mathcal{A}$ *makes at most* n *queries to* O *and outputs* $v \in \mathbb{S}^{d-1}$ *such that* $\mathbb{P}[\text{sign}(v \cdot x) \neq \text{sign}(u \cdot x)] \leq \varepsilon$, *then* $n \geq \frac{d \ln \frac{1}{\varepsilon}}{16(1-2\eta)^2}$.

Proof. We will prove this Lemma using Lemma 4.4.

First, we construct $\mathcal{W}$, $\mathcal{V}$, d, s, and P_θ. Let $\mathcal{W} = \mathbb{S}^{d-1}$. Let $\mathcal{V}$ be the set in Lemma 4.6 with $\gamma = 2\varepsilon$. For any $w_1, w_2 \in \mathcal{W}$, let $d(w_1, w_2) = \theta(w_1, w_2)$, $s = \pi\varepsilon$. Fix any algorithm $\mathcal{A}$. For any $w \in \mathcal{W}$, any $x \in \mathcal{X}$, define $P_w[Y = 1|X = x] = \begin{cases} 1-\eta, & w \cdot x \geq 0 \\ \eta, & w \cdot x < 0 \end{cases}$, and $P_w[Y = 0|X = x] = 1 - P_w[Y = 1|X = x]$. Define P_w^n to be the distribution of n examples $\left\{(X_i, Y_i)\right\}_{i=1}^n$ where Y_i is drawn from distribution $P_w(Y|X_i)$ and X_i is drawn by the active learning algorithm $\mathcal{A}$ based solely on the knowledge of $\left\{(X_j, Y_j)\right\}_{j=1}^{i-1}$.

By Lemma 4.6, we have $M = \left|\mathcal{V}\right| \geq \frac{\sqrt{d}}{2}\left(\frac{1}{4\pi\varepsilon}\right)^{d-1} - 1 \geq \frac{1}{4}\left(\frac{1}{4\pi\varepsilon}\right)^{d-1}$, and $d(w_1, w_2) \geq 2\pi\varepsilon = 2s$ for any distinct $w_1, w_2 \in \mathcal{V}$.

Clearly, for any $w \in \mathcal{W}$, if the optimal classifier is w, and the labeler O responds according to $P_w(\cdot \mid X = x)$, then it satisfies η-bounded noise condition. Therefore, to prove the lemma, it suffices to show that if $n \leq \frac{d \ln \frac{1}{\varepsilon}}{16(1-2\eta)^2}$, then

$$\sup_{w \in \mathcal{W}} P_w\left(d(w, \mathcal{A}(X^n, Y^n)) \geq s\right) \geq \frac{1}{2}.$$

Now, by Lemma 4.4,

$$\sup_{w\in\mathcal{W}} P_w^n\left(d(w,\mathcal{A}(X^n,Y^n))\geq s\right)\geq 1-\frac{I(V;X^n,Y^n)+\ln 2}{\ln M}\geq 1-\frac{I(V;X^n,Y^n)+\ln 2}{(d-1)\ln\frac{1}{4\pi\varepsilon}-\ln 4}.$$

It remains to show if $n=\frac{d\ln\frac{1}{\varepsilon}}{16(1-2\eta)^2}$, then $I(V;X^n,Y^n)\leq\frac{1}{2}\left((d-1)\ln\frac{1}{4\pi\varepsilon}-\ln 4\right)-\ln 2$.

By the chain rule of mutual information, we have

$$I(V;X^n,Y^n)=\sum_{i=1}^{n}\left(I\left(V;X_i\mid X^{i-1},Y^{i-1}\right)+I\left(V;Y_i\mid X^i,Y^{i-1}\right)\right)$$

First, we claim V and X_i are conditionally independent given $\left\{X^{i-1},Y^{i-1}\right\}$, and thus $I\left(V;X_i\mid X^{i-1},Y^{i-1}\right)=0$. The proof for this claim is as follows. Since the selection of X_i only depends on algorithm $\mathcal{A}$ and X^{i-1},Y^{i-1}, for any $v_1,v_2\in\mathcal{V}$, $\mathbb{P}\left(X_i\mid v_1,X^{i-1},Y^{i-1}\right)=\mathbb{P}\left(X_i\mid v_2,X^{i-1},Y^{i-1}\right)$. Thus,

$$\begin{aligned}\mathbb{P}\left(X_i\mid X^{i-1},Y^{i-1}\right) &= \sum_v\mathbb{P}\left(X_i,v\mid X^{i-1},Y^{i-1}\right)\\ &= \sum_v\mathbb{P}(v)\mathbb{P}\left(X_i\mid v,X^{i-1},Y^{i-1}\right)\\ &= \frac{1}{|\mathcal{V}|}\sum_v\mathbb{P}\left(X_i\mid v,X^{i-1},Y^{i-1}\right)\\ &= \mathbb{P}\left(X_i\mid V,X^{i-1},Y^{i-1}\right)\end{aligned}$$

Next, we show $I\left(V;Y_i\mid X^i,Y^{i-1}\right)\leq 5(1-2\eta)^2\ln 2$. On one hand, since $Y_i\in\{-1,+1\}$, $I\left(V;Y_i\mid X^i,Y^{i-1}\right)\leq H\left(V\mid X^i,Y^{i-1}\right)\leq\ln 2$. where $H(\cdot|\cdot)$ is the conditional entropy.

On the other hand,

$$\begin{aligned}
& I\left(V;Y_i \mid X^i,Y^{i-1}\right) \\
=& \mathbb{E}_{X^i,Y^i,V}\left[\ln \frac{\mathbb{P}\left(V,Y_i \mid X^i,Y^{i-1}\right)}{\mathbb{P}\left(V \mid X^i,Y^{i-1}\right)\mathbb{P}\left(Y_i \mid X^i,Y^{i-1}\right)}\right] \\
=& \mathbb{E}_{X^i,Y^i,V}\left[\ln \frac{\mathbb{P}\left(Y_i \mid V,X^i,Y^{i-1}\right)}{\mathbb{P}\left(Y_i \mid X^i,Y^{i-1}\right)}\right] \\
=& \mathbb{E}_{X^i,Y^i,V}\left[\ln \frac{\mathbb{P}\left(Y_i \mid V,X^i,Y^{i-1}\right)}{\mathbb{E}_{V'}\mathbb{P}\left(Y_i \mid V',X^i,Y^{i-1}\right)}\right] \\
\leq& \mathbb{E}_{X^i,Y^i,V,V'}\left[\ln \frac{\mathbb{P}\left(Y_i \mid V,X^i,Y^{i-1}\right)}{\mathbb{P}\left(Y_i \mid V',X^i,Y^{i-1}\right)}\right] \\
\leq& \max_{x^i,y^{i-1},v,v'} D_{\mathrm{KL}}\left(\mathbb{P}\left(Y_i \mid x^i,y^{i-1},v\right),\mathbb{P}\left(Y_i \mid x^i,y^{i-1},v'\right)\right) \\
=& \max_{x^i,y^{i-1},v,v'} D_{\mathrm{KL}}\left(\mathbb{P}\left(Y_i \mid x_i,v\right),\mathbb{P}\left(Y_i \mid x_i,v'\right)\right) \\
=& \max_{x^i,v,v'} D_{\mathrm{KL}}\left(P_v\left(Y_i \mid x_i\right),P_{v'}\left(Y_i \mid x'_i\right)\right) \\
\leq& \frac{(1-2\eta)^2}{\eta(1-\eta)}
\end{aligned}$$

where the first inequality follows from the convexity of KL-divergence, and the last inequality follows from Lemma 4.7.

Combining the two upper bounds, we get $I\left(V;Y_i \mid X^i,Y^{i-1}\right) \leq \min\left\{\ln 2, \frac{(1-2\eta)^2}{\eta(1-\eta)}\right\} \leq 5(1-2\eta)^2 \ln 2$.

Therefore, $I(V;X^n,Y^n) \leq 5n(1-2\eta)^2 \ln 2$. If $n \leq \frac{d \ln \frac{1}{\varepsilon}}{16(1-2\eta)^2} \leq \frac{\frac{1}{2}\left((d-1)\ln\frac{1}{4\pi\varepsilon}-\ln 4\right)-\ln 2}{5(1-2\eta)^2 \ln 2}$, then $I(V;X^n,Y^n) \leq \frac{1}{2}\left((d-1)\ln\frac{1}{4\pi\varepsilon}-\ln 4\right)-\ln 2$. This concludes the proof. □

Lemma 4.9. *For any $d > 0$, $0 \leq \eta < \frac{1}{2}$, $0 < \varepsilon < \frac{1}{3}$, $0 < \delta \leq \frac{1}{4}$, for any active learning algorithm $\mathcal{A}$, there is a $u \in \mathbb{S}^{d-1}$, and a labeler O that satisfies η-bounded noise condition with respect to u, such that if with probability at least $1-\delta$, $\mathcal{A}$ makes at most n queries to O and outputs $v \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(v \cdot x) \neq \text{sign}(u \cdot x)] \leq \varepsilon$, then $n \geq \Omega\left(\frac{\eta \ln \frac{1}{\delta}}{(1-2\eta)^2}\right)$.*

Proof. We prove this result by reducing the hypothesis testing problem in Lemma 4.5 to our problem of learning halfspaces.

Fix $d, \varepsilon, \delta, \eta$. Suppose $\mathcal{A}$ is an algorithm that for any $u \in \mathbb{S}^{d-1}$, under η-bounded noise condition, with probability at least $1-\delta$ outputs $v \in \mathbb{S}^{d-1}$ such that $\mathbb{P}[\text{sign}(v \cdot x) \neq \text{sign}(u \cdot x)] \leq \varepsilon < \frac{1}{3}$, which implies $\theta(v,u) \leq \frac{\pi}{3}$ under bounded noise condition.

Let $p_0 = \eta$, $p_1 = 1-\eta$. Suppose that $\alpha \sim$Bernoulli$(\frac{1}{2})$ is an unknown random variable. We are given a sequence of i.i.d. (given α) Bernoulli(p_α) random variables $\xi_1, \xi_2 \ldots$, and would like to test if α equals 0 or 1.

Define $e = (1,0,0,\ldots,0) \in \mathbb{R}^d$. Construct a labeler O such that for the i-th query x_i, it returns $2\xi_i - 1$ if $x_i \cdot e \geq 0$, and $1 - 2\xi_i$ otherwise. Clearly, the labeler O satisfies η-bounded noise condition with respect to underlying halfspace $u = (2\alpha-1)e = (2\alpha-1,0,0,\ldots,0) \in \mathbb{R}^d$.

Now, we run learning algorithm $\mathcal{A}$ with labeler O. Let m be the number of queries $\mathcal{A}$ makes, and $\mathcal{A}(\xi_1,\ldots,\xi_m)$ be the normal vector of the halfspace output by the learning algorithm. We define

$$f(\xi_1,\ldots,\xi_m) = \begin{cases} 0 & \text{if } \mathcal{A}(\xi_1,\ldots,\xi_m) \cdot e < 0 \\ 1 & \text{otherwise} \end{cases}.$$

By our assumption of $\mathcal{A}$ and construction of O, $\mathbb{P}\left(\theta\left(u, \mathcal{A}(\xi_1,\ldots,\xi_m)\right) \leq \frac{1}{3}\pi\right) \geq 1-\delta$, so $\mathbb{P}\left(f(\xi_1,\ldots,\xi_m) = \alpha\right) \geq 1-\delta$, implying $\mathbb{P}\left(f(\xi_1,\ldots,\xi_m) \neq \alpha\right) \leq \delta$. By Lemma 4.5, $m \geq$

$2\left\lfloor\frac{4\eta(1-\eta)}{(1-2\eta)^2}\ln\frac{1}{8\delta(1-2\delta)}\right\rfloor = \Omega\left(\frac{\eta\ln\frac{1}{\delta}}{(1-2\eta)^2}\right).$ □

4.4.2 Upper Bounds

We establish Theorem 4.10 in the bounded noise setting. The theorem implies that, with appropriate settings of input parameters, Algorithm 2 efficiently learns a halfspace of excess error at most ε with probability at least $1-\delta$, under the assumption that D_X is uniform over the unit sphere and O has bounded noise. In addition, it queries at most $\tilde{O}(\frac{d}{(1-2\eta)^2}\ln\frac{1}{\varepsilon})$ labels. This matches the lower bound in Theorem 4.3, and improves over the state of the art result of [ABHZ16], where a label complexity of $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})}\ln\frac{1}{\varepsilon})$ is shown using a different algorithm.

Theorem 4.10. *Suppose Algorithm 2 has inputs labeler O that satisfies η-bounded noise condition with respect to underlying halfspace u, initial halfspace v_0 such that $\theta(v_0,u)\leq\frac{\pi}{2}$, target error ε, confidence δ, sample schedule $\{m_k\}$ where $m_k = \lceil\frac{(3200\pi)^3 d}{(1-2\eta)^2}(\ln\frac{(3200\pi)^3 d}{(1-2\eta)^2}+\ln\frac{k(k+1)}{\delta})\rceil$, band width $\{b_k\}$ where $b_k = \frac{1}{2(600\pi)^2\ln\frac{m_k^2 k(k+1)}{\delta}}\frac{2^{-k}\pi(1-2\eta)}{\sqrt{d}}$. Then with probability at least $1-\delta$:*

1. *The output halfspace v is such that $\mathbb{P}[\mathrm{sign}(v\cdot X)\neq\mathrm{sign}(u\cdot X)]\leq\varepsilon$.*
2. *The number of label queries is $O\left(\frac{d}{(1-2\eta)^2}\cdot\ln\frac{1}{\varepsilon}\cdot\left(\ln\frac{d}{(1-2\eta)^2}+\ln\frac{1}{\delta}+\ln\ln\frac{1}{\varepsilon}\right)\right)$.*
3. *The number of unlabeled examples used is $O\left(\frac{d}{(1-2\eta)^3}\cdot\left(\ln\frac{d}{(1-2\eta)^2}+\ln\frac{1}{\delta}+\ln\ln\frac{1}{\varepsilon}\right)^2\cdot\frac{1}{\varepsilon}\ln\frac{1}{\varepsilon}\right)$.*
4. *The algorithm runs in time $O\left(\frac{d^2}{(1-2\eta)^3}\cdot\left(\ln\frac{d}{(1-2\eta)^2}+\ln\frac{1}{\delta}+\ln\ln\frac{1}{\varepsilon}\right)^2\cdot\frac{1}{\varepsilon}\ln\frac{1}{\varepsilon}\right)$.*

The theorem follows from Lemma 4.11 below. The key ingredient of the lemma is a delicate analysis of the dynamics of the angles $\{\theta_t\}_{t=0}^m$, where $\theta_t = \theta(w_t,u)$ is the angle between the iterate w_t and the halfspace u. Since x_t is randomly sampled and y_t is noisy, we are only able to show that θ_t decreases by a decent amount *in expectation*. To remedy the stochastic fluctuations, we apply martingale concentration inequalities to carefully control the upper envelope of sequence $\{\theta_t\}_{t=0}^m$.

Lemma 4.11. *Suppose Algorithm 3 has inputs labeler O that satisfies η-bounded noise condition with respect to underlying halfspace u, initial vector w_0 and angle upper bound $\theta \in (0, \frac{\pi}{2})$ such that $\theta(w_0, u) \le \theta$, confidence δ, number of iterations $m = \lceil \frac{(3200\pi)^3 d}{(1-2\eta)^2} (\ln \frac{(3200\pi)^3 d}{(1-2\eta)^2} + \ln \frac{1}{\delta}) \rceil$, band width $b = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{\delta}} \frac{\theta(1-2\eta)}{\sqrt{d}}$. then with probability at least $1-\delta$:*

1. *The output halfspace w_m is such that $\theta(w_m, u) \le \frac{\theta}{2}$.*
2. *The number of label queries is $O\left(\frac{d}{(1-2\eta)^2}\left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)\right)$.*
3. *The number of unlabeled examples drawn is $O\left(\frac{d}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.*
4. *The algorithm runs in time $O\left(\frac{d^2}{(1-2\eta)^3} \cdot \left(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{1}{\delta}\right)^2 \cdot \frac{1}{\theta}\right)$.*

In the rest of this subsection, we provide proofs for Lemma 4.11 and Theorem 4.10.

First, we give a generic lemma for the modified Perceptron update rule (4.1).

Lemma 4.12. *Suppose $w_t \in \mathbb{R}^d$ is a unit vector, and (x_t, y_t) is an labeled example where $x_t \in \mathbb{R}^d$ is a unit vector and $y_t \in \{-1, +1\}$. Let $\theta_t = \theta(u, w_t)$. Then, update*

$$w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t w_t \cdot x_t < 0\}(w_t \cdot x_t) \cdot x_t \tag{4.3}$$

gives an unit vector w_{t+1} such that

$$\cos\theta_{t+1} = \cos\theta_t - 2\mathbb{1}\{y_t w_t \cdot x_t < 0\}(w_t \cdot x_t) \cdot (u \cdot x_t) \tag{4.4}$$

Proof. We first show that w_{t+1} is still a unit vector. If $y_t = \text{sign}(w_t \cdot x_t)$, then $w_{t+1} = w_t$, thus it is still a unit vector; otherwise $w_{t+1} = w_t - 2(w_t \cdot x_t) \cdot x_t$. This gives that

$$\|w_{t+1}\|^2 = \|w_t\|^2 - 4(w_t \cdot x_t)(w_t \cdot x_t) + \|2(w_t \cdot x_t) \cdot x_t\|^2 = \|w_t\|^2 = 1.$$

This implies that $\cos\theta_t = w_t \cdot u$, and $\cos\theta_{t+1} = w_{t+1} \cdot u$. Now, taking inner products with u on both sides of Equation (4.3), we get

$$w_{t+1} \cdot u = w_t \cdot u - 2\mathbb{1}\left\{y_t w_t \cdot x_t < 0\right\}(w_t \cdot x_t) \cdot (u \cdot x_t)$$

which is equivalent to Equation (4.4). □

Next, we show that under the bounded noise model, $\cos\theta_t$ increases by a decent amount in expectation at each iteration of MODIFIED-PERCEPTRON (Algorithm 3), with appropriate settings of bandwidth b.

Lemma 4.13 (Progress Measure under Bounded Noise)**.** *Suppose* $0 < \tilde{c} < \frac{1}{288}$, $b = \frac{\tilde{c}(1-2\eta)\theta}{\sqrt{d}}$, $\theta \le \frac{27}{50}\pi$, *and* (x_t, y_t) *is drawn from* $D|_{R_t}$, *where* $R_t = \left\{(x,y) : x \cdot w_t \in [\frac{b}{2}, b]\right\}$. *Meanwhile, the labeler* O *satisfies the* η*-bounded noise condition. If unit vector* w_t *has angle* θ_t *with* u *such that* $\frac{1}{4}\theta \le \theta_t \le \frac{5}{3}\theta$, *then update* (4.3) *has the following guarantee:*

$$\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t\right] \ge \frac{\tilde{c}}{100\pi}\frac{(1-2\eta)^2\theta^2}{d}.$$

Proof. Define random variable $\xi = x_t \cdot w_t$. By the tower property of conditional expectation, $\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t\right] = \mathbb{E}\left[\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right] \mid \theta_t\right]$. Thus, it suffices to show

$$\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right] \ge \frac{\tilde{c}}{100\pi}\frac{(1-2\eta)^2\theta^2}{d}$$

for all $\theta_t \in [\frac{1}{4}\theta, \frac{5}{3}\theta]$ and $\xi \in [\frac{1}{2}b, b]$.

By Lemma 4.12, we know that

$$\cos\theta_{t+1} - \cos\theta_t = -2\mathbb{1}\left\{y_t \neq \mathrm{sign}(w_t \cdot x_t)\right\}(w_t \cdot x_t)\cdot(u \cdot x_t).$$

We simplify $\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right]$ as follows:

$$\begin{aligned}
& \mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right] \\
= \quad & \mathbb{E}\left[-2\xi u \cdot x_t \mathbb{1}\left\{y_t = -1\right\} \mid \theta_t, \xi\right] \\
= \quad & \mathbb{E}\left[-2\xi u \cdot x_t (\mathbb{1}\{u \cdot x_t > 0, y_t = -1\} + \mathbb{1}\{u \cdot x_t < 0, y_t = -1\}) \mid \theta_t, \xi\right] \\
\geq \quad & \mathbb{E}\left[-2\xi u \cdot x_t (\eta\mathbb{1}\{u \cdot x_t > 0\} + (1-\eta)\mathbb{1}\{u \cdot x_t < 0\}) \mid \theta_t, \xi\right] \\
= \quad & \mathbb{E}\left[-2\xi u \cdot x_t (\eta + (1-2\eta)\mathbb{1}\{u \cdot x_t < 0\}) \mid \theta_t, \xi\right] \\
= \quad & -2\xi\left(\eta\mathbb{E}\left[u \cdot x_t \mid \theta_t, \xi\right] + (1-2\eta)\mathbb{E}\left[u \cdot x_t \mathbb{1}\{u \cdot x_t < 0\} \mid \theta_t, \xi\right]\right) \qquad (4.5)
\end{aligned}$$

where the second equality is from algebra, the first inequality is from that $\mathbb{P}[y_t = -1 | u \cdot x_t > 0] \leq \eta$ and $\mathbb{P}[y_t = -1 | u \cdot x_t < 0] \geq 1 - \eta$, the last two equalities are from algebra.

By Lemma A.9 and that $0 \leq \theta_t \leq \frac{5}{3}\theta \leq \frac{9}{10}\pi$, we have $\mathbb{E}[u \cdot x_t \mathbb{1}\left\{u \cdot x_t < 0\right\} | \theta_t, \xi] \leq \xi - \frac{\theta_t}{36\sqrt{d}}$, and $\mathbb{E}[u \cdot x_t | \theta_t, \xi] \leq \xi$.

Thus,

$$
\begin{aligned}
& \mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right] \\
\geq\ & -2\xi(\xi\eta + (\xi - \frac{\theta_t}{36\sqrt{d}})(1-2\eta)) \\
\geq\ & 2\xi(\frac{\theta_t}{36\sqrt{d}}(1-2\eta) - \xi) \\
\geq\ & b\frac{\theta_t}{72\sqrt{d}}(1-2\eta) \\
\geq\ & \frac{\tilde{c}}{100\pi}\frac{(1-2\eta)^2\theta^2}{d}
\end{aligned}
$$

where the first and second inequalities are from algebra, the third inequality is from that $\xi \leq b \leq \frac{\theta(1-2\eta)}{288\sqrt{d}} \leq \frac{\theta_t(1-2\eta)}{72\sqrt{d}}$, and that $\xi \geq \frac{b}{2}$. the last inequality is by expanding $b = \frac{\tilde{c}(1-2\eta)\theta}{\sqrt{d}}$ and that $\theta_t \geq \frac{\theta}{4}$.

In conclusion, if $\frac{1}{4}\theta \leq \theta_t \leq \frac{5}{3}\theta$, then $\mathbb{E}\left[\cos\theta_{t+1} - \cos\theta_t \mid \theta_t, \xi\right] \geq \frac{\tilde{c}}{100\pi}\frac{(1-2\eta)^2\theta^2}{d}$ for $\xi \in [\frac{b}{2}, b]$. The lemma follows. □

Next, we present two major building blocks of Lemma 4.11.

The first building block is a technical lemma that coarsely bounds the difference between $\cos\theta_{t+1}$ and $\cos\theta_t$.

Lemma 4.14. *Suppose* $0 < \tilde{c}, \zeta < 1$, $b = \frac{\tilde{c}\zeta\theta}{\sqrt{d}} \leq 1$, *and* (x_t, y_t) *is drawn from distribution* $D|_{R_t}$ *where* $R_t = \left\{(x,y) : x \cdot w_t \in [\frac{b}{2}, b]\right\}$. *If unit vector* w_t *has angle* θ_t *with* u *such that* $\theta_t \leq \frac{5}{3}\theta$, *then update* (4.3) *has the following guarantee:* $|\cos\theta_{t+1} - \cos\theta_t| \leq \frac{16\tilde{c}\zeta\theta^2}{3\sqrt{d}}$.

Proof. By Lemma 4.12,

$$\cos\theta_{t+1} - \cos\theta_t = -2\mathbb{1}\left\{y_t \neq \text{sign}(w_t \cdot x_t)\right\}(w_t \cdot x_t) \cdot (u \cdot x_t).$$

Firstly, note $|\cos\theta_{t+1} - \cos\theta_t| \le 2\,|w_t \cdot x_t|\,|u \cdot x_t| \le 2b\,|u \cdot x_t|$.

Observe that

$$\begin{aligned} & |u \cdot x_t| \\ \le\; & |w_t \cdot x_t| + \left|(u - w_t) \cdot x_t\right| \\ \le\; & b + 2\sin\frac{\theta_t}{2} \\ \le\; & b + \theta_t \end{aligned}$$

Thus, we have $|\cos\theta_{t+1} - \cos\theta_t| \le 2b(b+\theta_t) = \frac{2\tilde{c}^2\zeta^2\theta^2}{d} + \frac{2\tilde{c}\zeta\theta\theta_t}{\sqrt{d}} \le \frac{16\tilde{c}\zeta\theta^2}{3\sqrt{d}}$. □

The second building block is a lemma that turns per-iteration in-expectation guarantees provided by Lemma 4.13 into high probability upper bounds on the final θ_m.

Lemma 4.15. *Suppose* $0 < \zeta < 1$, *and the following conditions hold:*

1. *Initial unit vector* w_0 *has angle* $\theta_0 = \theta(w_0, u) \le \theta \le \frac{27}{50}\pi$ *with* u*;*
2. *Integer* $m = \lceil \frac{(3200\pi)^3 d}{\zeta^2}(\ln \frac{(3200\pi)^3 d}{\zeta^2} + \ln\frac{1}{\delta})\rceil$ *and* $\tilde{c} = \frac{1}{2(600\pi)^2 \ln\frac{m^2}{\delta}}$*;*
3. *For all* t*, if* $\frac{1}{4}\theta \le \theta_t \le \frac{5}{3}\theta$*, then* $\mathbb{E}[\cos\theta_{t+1} - \cos\theta_t | \theta_t] \ge \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}$*;*
4. *For all* t*, if* $\theta_t \le \frac{5}{3}\theta$*, then* $|\cos\theta_{t+1} - \cos\theta_t| \le \frac{16\tilde{c}\zeta\theta^2}{3\sqrt{d}}$ *holds with probability 1.*

Then with probability at least $1 - \delta/2$*,* $\theta_m \le \frac{1}{2}\theta$.

Proof. Define random variable D_t as:

$$D_t := \left(\cos\theta_{t+1} - \cos\theta_t - \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}\right)\mathbb{1}\left\{\frac{1}{4}\theta \le \theta_t \le \frac{5}{3}\theta\right\}$$

Note that $\mathbb{E}[D_t|\theta_t] \geq 0$ and from Lemma 4.14, $|D_t| \leq |\cos\theta_{t+1} - \cos\theta_t| + \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d} \leq \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}}$. Therefore, $\{D_t\}$ is a bounded submartingale difference sequence. By Azuma's Inequality (see Lemma A.5) and union bound, define event

$$E = \left\{ \text{for all } 0 \leq t_1 \leq t_2 \leq m, \sum_{s=t_1}^{t_2-1} D_s \geq -\frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}}\sqrt{2(t_2-t_1)\ln\frac{2m^2}{\delta}} \right\}$$

Then $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$.

We now condition on event E. We break the subsequent analysis into two parts: (1) Show that there exists some t such that θ_t goes below $\frac{1}{4}\theta$. (2) Show that θ_t must stay below $\frac{1}{2}\theta$ afterwards.

1. First, it can be checked by algebra that $m \geq \frac{200\pi d}{\zeta^2\tilde{c}}$. We show the following claim.

   Claim 4.16. *There exists some $t \in [0,m]$, such that $\theta_t < \frac{1}{4}\theta$.*

   Proof. We first show that it is impossible for all $t \in [0,m]$ such that $\theta_t \in \left[\frac{1}{4}\theta, \frac{5}{3}\theta\right]$. To this end, assume this holds for the sake of contradiction. In this case, for all $t \in [0,m]$, $D_t = \cos\theta_{t+1} - \cos\theta_t - \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}$. Therefore,

$$\begin{aligned}
& \cos\theta_m - \cos\theta_0 \\
= & \sum_{s=0}^{m-1} D_s + \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}m \\
\geq & \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}m - \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}}\sqrt{2m\ln\frac{m^2}{\delta}} \\
\geq & \frac{\theta^2}{100\pi}\left[\frac{\tilde{c}\zeta^2 m}{d} - \sqrt{\frac{\tilde{c}\zeta^2 m}{d}}\right] \\
\geq & \theta^2
\end{aligned}$$

   where the first inequality is from the definition of event E, the second inequality is from that $\tilde{c} = \frac{1}{2(600\pi)^2 \ln \frac{m^2}{\delta}}$, the third inequality is from that $\frac{\tilde{c}\zeta^2 m}{d} \geq 200\pi$.

Since $\cos\theta_0 \geq \cos\theta \geq 1 - \frac{1}{2}\theta^2$, this gives that $\cos\theta_m \geq 1 + \frac{1}{2}\theta^2 > 1$, contradiction.

Next, define $\tau := \min\left\{ t \geq 0 : \theta_t \notin \left[\frac{1}{4}\theta, \frac{5}{3}\theta\right] \right\}$. We now know that $\tau \leq m$ by the reasoning above. It suffices to show that $\theta_\tau < \frac{1}{4}\theta$, that is, the first time when θ_t goes outside the interval $[\frac{1}{4}\theta, \frac{5}{3}\theta]$, it must be crossing the left boundary as opposed to the right one.

By the definition of τ, for all $0 \leq t \leq \tau - 1$, $\theta_\tau \in \left[\frac{1}{4}\theta, \frac{5}{3}\theta\right]$. Thus,

$$\begin{aligned} & \cos\theta_\tau - \cos\theta_0 \\ = \quad & \sum_{t=0}^{\tau-1} D_t + \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}\tau \\ \geq \quad & \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}\tau - \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}}\sqrt{\tau \ln\frac{m^2}{\delta}} \\ \geq \quad & -900\pi \ln\frac{m^2}{\delta}\tilde{c}\theta^2 \geq -\frac{1}{75}\theta^2 \end{aligned} \tag{4.6}$$

where the first inequality is by the definition of E; the second inequality is by minimizing over $\tau \in [0, m]$; the last inequality is from the definition of $\tilde{c}$.

Now, if $\theta_\tau \geq \frac{5}{3}\theta$, then

$$\begin{aligned} \cos\theta_\tau - \cos\theta_0 \quad & \leq \quad \cos\frac{5}{3}\theta - \cos\theta \\ & \leq \quad 1 - \frac{1}{5}\left(\frac{5}{3}\right)^2\theta^2 - 1 + \frac{1}{2}\theta^2 \\ & < \quad -\frac{1}{75}\theta^2 \end{aligned}$$

where the first inequality follows from $\theta_\tau \geq \frac{5}{3}\theta$ and $\theta_0 \leq \theta$, and the second inequality follows from Lemma A.3. This contradicts with Inequality (4.6).

This gives that $\theta_\tau < \frac{5}{3}\theta$. Since $\theta_\tau \notin \left[\frac{1}{4}\theta, \frac{5}{3}\theta\right]$, it must be the case that $\theta_\tau < \frac{1}{4}\theta$. □

2. We now show the following claim to conclude the proof.

Claim 4.17. *θ_m, the angle in the last iteration, is at most $\frac{1}{2}\theta$.*

Proof. Define $\sigma = \max\left\{t \in [0,m] : \theta_t < \frac{1}{4}\theta\right\}$. by Claim 4.16, such σ is well-defined on event E. We now show that θ_t will not exceed $\frac{1}{2}\theta$ afterwards. Assume for the sake of contradiction that for some $t > \sigma$, $\theta_t > \frac{1}{2}\theta$.

Now define $\gamma := \min\left\{t > \sigma : \theta_t > \frac{1}{2}\theta\right\}$. We know by the definitions of σ and γ, for all $t \in [\sigma+1, \gamma-1]$, $\theta_t \in [\frac{1}{4}\theta, \frac{1}{2}\theta]$. Thus,

$$\begin{aligned}
& \cos\theta_\gamma - \cos\theta_{\sigma+1} \\
= \quad & \sum_{t=\sigma+1}^{\gamma-1} D_t + \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}(\gamma-\sigma-1) \\
\geq \quad & \frac{\tilde{c}}{100\pi}\frac{\zeta^2\theta^2}{d}(\gamma-\sigma-1) - \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}}\sqrt{(\gamma-\sigma-1)\ln\frac{m^2}{\delta}} \\
\geq \quad & -900\pi\ln\frac{m^2}{\delta}\tilde{c} \geq -\frac{1}{75}\theta^2
\end{aligned} \tag{4.7}$$

where the first inequality is by the definition of E; the second inequality is by minimization over $\gamma-\sigma-1 \in [0,m]$; the last inequality is from the definition of $\tilde{c}$.

On the other hand, $\theta_\gamma > \frac{1}{2}\theta$ and $\theta_\sigma < \frac{1}{4}\theta$. We have

$$\begin{aligned}
\cos\theta_\gamma - \cos\theta_{\sigma+1} \quad & \leq \quad \cos\theta_\gamma - \cos\theta_\sigma + \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}} \\
& \leq \quad \cos\frac{\theta}{2} - \cos\frac{\theta}{4} + \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}} \\
& \leq \quad 1 - \frac{1}{20}\theta^2 - 1 + \frac{1}{32}\theta^2 + \frac{6\tilde{c}\zeta\theta^2}{\sqrt{d}} \\
& < \quad -\frac{1}{75}\theta^2
\end{aligned}$$

where the first inequality follows from Lemma 4.14, the third follows from Lemma A.3, and the last follows from algebra. This contradicts with Inequality (4.7). □

Thus, with probability at least $1-\delta/2$, $\theta_m \leq \frac{1}{2}\theta$. □

Now, we are ready to present the proofs of Lemma 4.11 and Theorem 4.10.

Proof of Lemma 4.11. We show that each item holds with high probability respectively.

1. It can be verified that conditions for Lemma 4.15 are satisfied with $\zeta = 1-2\eta$ (item 3 in the condition follows from Lemma 4.13, and item 4 in the condition follows from Lemma 4.14). This shows that items 1 with probability at least $1-\delta/2$.
2. By the definition of m, the number of label queries is $m = O\left(\frac{d}{(1-2\eta)^2}\log\frac{d}{\delta(1-2\eta)^2}\right)$.
3. As for the number of unlabeled examples drawn by the algorithm, at each iteration $t \in [0,m]$, it takes Z_t trials to hit an example in $[\frac{b}{2}, b]$, where Z_t is a Geometric(p) random variable with $p = \mathbb{P}_{x\sim D_X}[w_t \cdot x \in [\frac{b}{2}, b]]$. From Lemma A.8, $p \geq \frac{\sqrt{d}}{8\pi}b = \frac{\tilde{c}(1-2\eta)\theta}{8\pi} = \Omega(\frac{(1-2\eta)\theta}{\ln\frac{d}{\delta(1-2\eta)^2}})$.

   Define event
   $$E := \left\{Z_1 + \ldots + Z_m \leq \frac{2m}{p}\right\}$$
   From Lemma A.6 and the choice of m, $\mathbb{P}[E] \geq 1-\frac{\delta}{2}$. Thus, on event E, the total number of unlabeled examples drawn is at most $\frac{2m}{p} = O(\frac{d}{(1-2\eta)^3}\log^2\frac{d}{\delta(1-2\eta)^2}\frac{1}{\theta})$.
4. Observe that the time complexity for processing each example is at most $O(d)$. This shows that on event E, the total running time of the algorithm is at most $O(d\cdot\frac{2m}{p}) = O(\frac{d^2}{(1-2\eta)^3}\log^2\frac{d}{\delta(1-2\eta)^2}\frac{1}{\theta})$.

Therefore, by a union bound, with probability at least $1-\delta$, items 1 to 4 hold simultaneously. □

Proof of Theorem 4.10. From Lemma 4.11, we know that for every k, there is an event E_k such that $\mathbb{P}(E_k) \geq 1 - \frac{\delta}{k(k+1)}$, and on event E_k, items 1 to 4 of Lemma 4.11 hold for input $w_0 = v_{k-1}$, output $w_m = v_k$, $\theta = \frac{\pi}{2^k}$, $\delta = \frac{\delta}{k(k+1)}$.

Define event $E = \cup_{k=1}^{k_0} E_k$. By union bound, $\mathbb{P}(E) \geq 1 - \delta$. We henceforth condition on event E happening.

1. By induction, the final output $v = v_{k_0}$ is such that $\theta(v,u) \leq 2^{-k_0}\pi \leq \varepsilon\pi$, implying that $\mathbb{P}[\text{sign}(v \cdot X) \neq \text{sign}(u \cdot X)] \leq \varepsilon$.

2. Define the number of label queries to labeler O at iteration k as m_k. On event E_k, m_k is at most $O\left(\frac{d}{(1-2\eta)^2}\left(\ln\frac{d}{(1-2\eta)^2} + \ln\frac{k}{\delta}\right)\right)$. Thus, the total number of label queries to labeler O is $\sum_{k=1}^{k_0} m_k$, which is at most

$$k_0 \cdot m_{k_0} = O\left(k_0 \cdot \frac{d}{(1-2\eta)^2}\left(\ln\frac{d}{(1-2\eta)^2} + \ln\frac{k_0}{\delta}\right)\right).$$

   Item 2 is proved by noting that $k_0 \leq \log\frac{1}{\varepsilon} + 1$.

3. Define the number of unlabeled examples drawn iteration k as n_k. On event E_k, n_k is at most $O\left(\frac{d}{(1-2\eta)^3} \cdot \left(\ln\frac{d}{(1-2\eta)^2} + \ln\frac{k}{\delta}\right)^2 \cdot \frac{1}{\varepsilon}\right)$. Thus, the total number of unlabeled examples drawn is $\sum_{k=1}^{k_0} n_k$, which is at most

$$k_0 n_{k_0} = O\left(k_0 \cdot \frac{d}{(1-2\eta)^3} \cdot \left(\ln\frac{d}{(1-2\eta)^2} + \ln\frac{k_0}{\delta}\right)^2 \cdot \frac{1}{\varepsilon}\right).$$

   Item 3 is proved by noting that $k_0 \leq \log\frac{1}{\varepsilon} + 1$.

4. Item 4 is immediate from Item 3 and the fact that the time for processing each example is at most $O(d)$. □

Table 4.1: A comparison of algorithms for active learning of halfspaces under the uniform distribution, in the η-bounded noise model.

Algorithm	Label Complexity	Time Complexity
[BBZ07, BL13, ZC14]	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	$\text{superpoly}(d, \frac{1}{\epsilon})$ [1]
[ABHZ16]	$\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \cdot \ln \frac{1}{\epsilon})$	$\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \cdot \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{(1-2\eta)^2} \ln \frac{1}{\epsilon})$	$\tilde{O}\left(\frac{d^2}{(1-2\eta)^3} \frac{1}{\epsilon}\right)$

4.5 Discussion

4.5.1 Comparisons

We have shown that in the η-*bounded noise setting*, the proposed Algorithm 2 runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3\epsilon}\right)$, and requires $\tilde{O}\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$ labels. This label complexity almost matches the information-theoretic lower bound of $\Omega\left(\frac{d}{(1-2\eta)^2} \cdot \ln \frac{1}{\epsilon}\right)$, and thus is *nearly optimal*. Our time and label complexities substantially improve over the state of the art result of [ABHZ16], which runs in time $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \frac{1}{\epsilon})$ and requires $\tilde{O}(d^{O(\frac{1}{(1-2\eta)^4})} \ln \frac{1}{\epsilon})$ labels.

Table 4.1 presents comparisons between our results and results most closely related to ours.

In our algorithm and analysis, we assume the unlabeled examples are drawn uniformly from the unit sphere. However, they can be easily generalized to any spherical symmetrical distributions, for example, isotropic Gaussian distributions. They can also be generalized to distributions whose densities with respect to uniform distribution are bounded away from 0.

[1] The algorithm needs to minimize 0-1 loss, the best known method for which requires superpolynomial time.

4.5.2 Implications to Passive Learning

Algorithm 2 can be converted to a passive learning algorithm, Algorithm 4, for learning homogeneous halfspaces under the uniform distribution over the unit sphere. Algorithm 4 has PAC sample complexities close to the lower bounds under bounded noise.

The algorithmic framework is similar to Algorithm 2, except that it calls Algorithm 5 rather than Algorithm 3.

Algorithm 4 PASSIVE-PERCEPTRON

```
Input: Initial halfspace $v_0$, target error $\epsilon$, confidence $\delta$, sample schedule $\{m_k\}$, band width $\{b_k\}$.
Output: learned halfspace $\hat{v}$.
Let $k_0 = \lceil \log_2 \frac{1}{\epsilon} \rceil$.
for $k = 1, 2, \ldots, k_0$ do
    $v_k \leftarrow$ PASSIVE-MODIFIED-PERCEPTRON$(O, v_{k-1}, \frac{\pi}{2^k}, \frac{\delta}{k(k+1)}, m_k, b_k)$.
end for
Return $v_{k_0}$.
```

Algorithm 5 is similar to Algorithm 3, except that it draws labeled examples from D directly, as opposed to performing label queries on unlabeled examples drawn.

It can be seen that with the same input as Algorithm 2, Algorithm 4 has exactly the same running time, and the number of labeled examples drawn in Algorithm 4 is exactly the same as the number of unlabeled examples drawn in Algorithm 2. We have the following corollary which is the immediate consequence of Theorem 4.10.

Corollary 4.18 (PASSIVE-PERCEPTRON under Bounded Noise)**.** *Suppose Algorithm 4 has inputs distribution D that satisfies η-bounded noise condition with respect to u, initial halfspace v_0, target error ϵ, confidence δ, sample schedule $\{m_k\}$ where $m_k = \Theta\left(\frac{d}{(1-2\eta)^2}(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta})\right)$,*

Algorithm 5 PASSIVE-MODIFIED-PERCEPTRON

```
Input: Initial halfspace $w_0$, angle upper bound $\theta$, confidence $\delta$, number of iterations $m$, band width $b$.
Output: Improved halfspace $w_m$.
for $t = 0, 1, 2, \ldots, m-1$ do
    Define region $C_t = \left\{(x,y) \in \mathbb{S}^{d-1} \times \{-1,+1\} : \frac{b}{2} \le w_t \cdot x \le b\right\}$.
    Rejection sample $(x_t, y_t) \sim D|_{C_t}$. In other words, repeat drawing example $(x_t, y_t) \sim D$ until it is in $C_t$.
    $w_{t+1} \leftarrow w_t - 2\mathbb{1}\{y_t w_t \cdot x_t < 0\} \cdot (w_t \cdot x_t) \cdot x_t$.
end for
Return $w_m$.
```

Table 4.2: A comparison of algorithms for PAC learning halfspaces under the uniform distribution, in the η-bounded noise model.

Algorithm	Sample Complexity	Time Complexity
[ABHZ16]	$\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$	$\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$
ERM [MN06]	$\tilde{O}(\frac{d}{(1-2\eta)\epsilon})$	superpoly$(d, \frac{1}{\epsilon})$
Our Work	$\tilde{O}(\frac{d}{(1-2\eta)^3\epsilon})$	$\tilde{O}(\frac{d^2}{(1-2\eta)^3} \cdot \frac{1}{\epsilon})$

band width $\{b_k\}$ where $b_k = \Theta\left(\frac{2^{-k}(1-2\eta)}{\sqrt{d}\ln(km_k/\delta)}\right)$. Then with probability at least $1-\delta$: (1) The output halfspace v is such that $l(h_v) \le l(h_u) + \epsilon$; (2) The number of labeled examples drawn is $\tilde{O}\left(\frac{d}{(1-2\eta)^3\epsilon}\right)$. (3) The algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3\epsilon}\right)$.

In the η-bounded noise model, the sample complexity of PASSIVE-PERCEPTRON improves over the state of the art result of [ABHZ16], where a sample complexity of $\tilde{O}(\frac{d^{O(\frac{1}{(1-2\eta)^4})}}{\epsilon})$ is obtained. The bound has the same dependency on ϵ and d as the minimax upper bound of $\tilde{\Theta}(\frac{d}{\epsilon(1-2\eta)})$ by [MN06], which is achieved by a computationally inefficient ERM algorithm.

Table 4.2 presents comparisons between our results and results most closely related to ours.

4.6 Acknowledgements

This chapter is based on the material as it appears in Advances in Neural Information Processing Systems 2017 (Songbai Yan and Chicheng Zhang, “Revisiting perceptron: Efficient and label-optimal active learning of halfspaces”) [YZ17]. The dissertation author is the co-primary investigator and co-author of this material.

Chapter 5

Active Learning with Abstention Feedback

5.1 Introduction

In this chapter, we consider a new interactive model for active learning, where in addition to providing a possibly noisy label, the labeler can sometimes abstain from labeling. This scenario arises naturally in difficult labeling tasks and has been considered in computer vision by [FZ12, KFR$^+$15]. Our goal in this chapter is to investigate this problem from a foundational perspective, and explore what kind of conditions are needed, and how an abstaining labeler can affect properties such as consistency and query complexity of active learning algorithms.

We first consider a condition where the probability that the labeler abstains is upper bounded by a monotonic function, so that the labeler can abstain with a higher probability as the instance being queried is closer to the decision boundary. We provide an information-theoretic query complexity lower bound for any active learning algorithms and an algorithm with a query complexity bound that almost matches the lower bound. This nearly-optimal algorithm, however, simply ignores abstention feedback, suggesting that in order to enable the learner to utilize abstention feedback, the labeler needs to satisfy stronger conditions.

Consequently, we consider a stronger condition where the probability that the labeler abstains increases strictly monotonically close to the decision boundary. We propose an active learning algorithm that is capable of exploiting this condition. We also prove that this algorithm achieves nearly optimal query complexity bounds. An important property of this algorithm is that the improvement of query complexity is achieved in a *completely adaptive manner*: it needs *no information whatsoever on the abstention rates or rates of label noise*. This algorithm is statistically consistent under a very mild condition — when the abstention rate is non-decreasing as we get closer to the decision boundary. Under a slightly stronger additional condition where the abstention rate is upper-bounded, this algorithm has the same query complexity as our former algorithm. However, if the abstention rate of the labeler increases strictly monotonically close to the decision boundary, then this algorithm adapts and does substantially better: it simply exploits the increasing abstention rate close to the decision boundary, and does not even have to rely on the noisy labels! Our result also strengthens existing results on active learning from (non-abstaining) noisy labelers by providing an adaptive algorithm that achieves that same performance as [CN08] without knowledge of noise parameters.

5.2 Setup

We consider active learning for binary classification. We are given an instance space $\mathcal{X} = [0,1]^d$ and a label space $\mathcal{L} = \{0,1\}$. Each instance $x \in \mathcal{X}$ is assigned to a label $l \in \{0,1\}$ by an underlying function $h^* : \mathcal{X} \to \{0,1\}$ unknown to the learning algorithm in a hypothesis space $\mathcal{H}$ of interest. The learning algorithm has access to any $x \in \mathcal{X}$, but no access to their labels. Instead, it can only obtain label information through interactions with a labeler, whose relation to h^* is to be specified later. The objective of the algorithm is to sequentially select the instances to query for label information and output a classifier $\hat{h}$ that is close to h^* while making as few queries as possible.

We consider a non-parametric setting as in [CN08, Min12] where the hypothesis space $\mathcal{H} = \{h_g(x) = \mathbb{1}\left[x_d > g(\tilde{x})\right] \mid g : [0,1]^{d-1} \to [0,1] \text{ is } (K,\gamma)\text{-Hölder smooth}\}$ is the *smooth boundary fragment* class (recall $\tilde{x} \in \mathbb{R}^{d-1}$ is the first $d-1$ coordinates of the vector x). In other words, the decision boundaries of classifiers in this class are epigraph of smooth functions (see Figure 5.1 for example). We assume $h^*(x) = \mathbb{1}\left[x_d > g^*(\tilde{x})\right] \in \mathcal{H}$. When $d = 1$, $\mathcal{H}$ reduces to the space of threshold functions $\{h_\theta(x) = \mathbb{1}[x > \theta] : \theta \in [0,1]\}$.

The performance of a classifier $h(x) = \mathbb{1}\left[x_d > g(\tilde{x})\right]$ is evaluated by the L^1 distance between the decision boundaries $\|g - g^*\| = \int_{[0,1]^{d-1}} \left|g(\tilde{x}) - g^*(\tilde{x})\right| \mathrm{d}\tilde{x}$.

The learning algorithm can only obtain label information by querying a labeler who is allowed to abstain from labeling or return an incorrect label (flipping between 0 and 1). For each query $x \in [0,1]^d$, the labeler L will return $y \in \mathcal{Y} = \{0, 1, \perp\}$ ($\perp$ means that the labeler abstains from providing a 0/1 label) according to some distribution $P_L(Y = y \mid X = x)$. When it is clear from the context, we will drop the subscript from $P_L(Y \mid X)$. Note that while the labeler can declare its indecision by outputting $\perp$, we do not allow classifiers in our hypothesis space to output $\perp$.

In our active learning setting, our goal is to output a boundary g that is close to g^* while making as few interactive queries to the labeler as possible. In particular, we want to find an algorithm with low *query complexity* $\Lambda(\varepsilon, \delta, \mathcal{A}, L, g^*)$, which is defined as the minimum number of queries that Algorithm $\mathcal{A}$, acting on samples with ground truth g^*, should make to a labeler L to ensure that the output classifier $h_g(x) = \mathbb{1}\left[x_d > g(\tilde{x})\right]$ has the property $\|g - g^*\| = \int_{[0,1]^{d-1}} \left|g(\tilde{x}) - g^*(\tilde{x})\right| d\tilde{x} \leq \varepsilon$ with probability at least $1 - \delta$ over the responses of L.

5.2.1 Conditions for the Labeler

We now introduce three conditions on the response of the labeler.

Condition 1. The response distribution of the labeler $P(Y \mid X)$ satisfies:

- (abstention) For any $\tilde{x} \in [0,1]^{d-1}$, $x_d, x'_d \in [0,1]$, if $\left|x_d - g^*(\tilde{x})\right| \geq \left|x'_d - g^*(\tilde{x})\right|$ then $P(\perp \mid (\tilde{x}, x_d)) \leq P(\perp \mid (\tilde{x}, x'_d))$;
- (noise) For any $x \in [0,1]^d$, $P(Y \neq \mathbb{1}\left[x_d > g^*(\tilde{x})\right] \mid x, Y \neq \perp) \leq \frac{1}{2}$.

Condition 1 means that the closer the instance x is to the decision boundary $\left(\tilde{x}, g^*(\tilde{x})\right)$, the more likely the labeler is to abstain from labeling. This complies with the intuition that instances closer to the decision boundary are harder to classify. The 0/1 labels can be flipped with probability as large as $\frac{1}{2}$. In other words, we allow unbounded noise.

Condition 2. Let C, β be non-negative constants, and $f : [0,1] \to [0,1]$ be a nondecreasing function. The response distribution $P(Y \mid X)$ satisfies:

- (abstention) $P(\perp \mid x) \leq 1 - f\left(\left|x_d - g^*(\tilde{x})\right|\right)$;
- (noise) $P(Y \neq \mathbb{1}\left[x_d > g^*(\tilde{x})\right] \mid x, Y \neq \perp) \leq \frac{1}{2}\left(1 - C\left|x_d - g^*(\tilde{x})\right|^{\beta}\right)$.

Condition 2 requires the abstention and noise probabilities to be upper-bounded, and these upper bounds decrease as x moves further away from the decision boundary. The abstention rate can be 1 at the decision boundary, so the labeler may always abstain at the decision boundary. The condition on the noise satisfies the popular Tsybakov noise condition [Tsy04], and a similar condition was considered by [CN08].

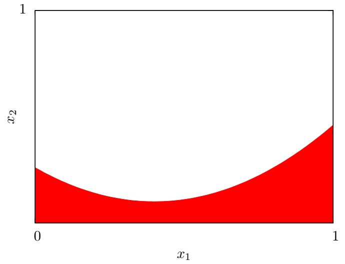


Figure 5.1: A classifier with boundary $g(\tilde{x}) = (x_1 - 0.4)^2 + 0.1$ for $d = 2$. Label 1 is assigned to the region above, 0 to the below (red region)

Condition 3. Let $f : [0,1] \to [0,1]$ be a nondecreasing function such that $\exists 0 < c < 1, \forall 0 < a \leq 1$ $\forall 0 \leq b \leq \frac{2}{3}a$, $\frac{f(b)}{f(a)} \leq 1 - c$. The response distribution satisfies: $P(\perp | \, x) = 1 - f\left(\left|x_d - g^*(\tilde{x})\right|\right)$.

An example where Condition 3 holds is $P(\perp | \, x) = 1 - (x - 0.3)^{\alpha}$ ($\alpha > 0$).

Condition 3 requires the abstention probability $P(\perp | (\tilde{x}, x_d))$ to be not too flat with respect to x_d. For example, when $d = 1$, $P(\perp | \, x) = 0.68$ for $0.2 \leq x \leq 0.4$ (shown as Figure 5.2 (left)) does not satisfy Condition 3, and abstention responses are not informative since this abstention rate alone yields no information on the location of the decision boundary. In contrast, $P(\perp | \, x) = 1 - \sqrt{|x - 0.3|}$ (shown as Figure 5.2 (right)) satisfies Condition 3, and the learner could infer it is getting close to the decision boundary when it starts receiving more abstention responses.

Note that here c, f, C, β are parameters that characterize the complexity of the learning task. We want to design an algorithm that does not require knowledge of these parameters and still achieves nearly optimal query complexity.

In the following two sections, we consider the one-dimensional case ($d = 1$) to demonstrate the main idea. We extend the discussion to the d-dimensional instance space in the last section of this chapter.

When $d = 1$, the decision boundary g^* becomes a point in $[0,1]$, and the corresponding classifier is a threshold function over [0,1]. In other words the hypothesis space becomes

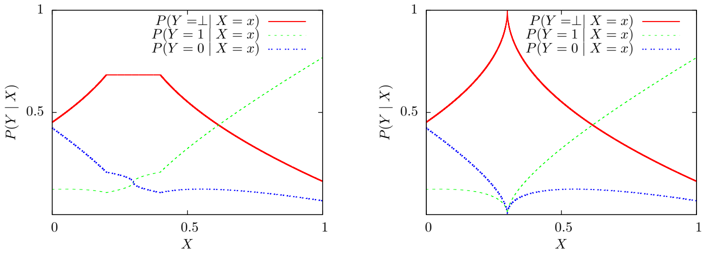


Figure 5.2: Left: The distributions satisfies Conditions 1 and 2, but the abstention feedback is useless since $P(\perp | x)$ is flat between $x = 0.2$ and 0.4. Right: The distributions satisfies Conditions 1, 2, and 3. The abstention feedback can be used to save queries.

$\mathcal{H} = \{f_\theta(x) = \mathbb{1}[x > \theta] : \theta \in [0,1]\})$. We denote the ground truth decision boundary by $\theta^* \in [0,1]$. We want to find a $\hat{\theta} \in [0,1]$ such that $|\hat{\theta} - \theta^*|$ is small while making as few queries as possible.

5.3 Active Learning with Flat Abstention Rates

In this section, we consider active learning under Condition 2 where the abstention probability of the labeler is upper-bounded by some monotonic function but can be flat. We first derive an information-theoretic query complexity lower bound for any active learning algorithms. Then, we provide an algorithm that simply ignores abstention feedback while achieving a query complexity upper bound that almost matches the lower bound. This implies that if we would like to improve the query complexity of the algorithm by making use of abstention feedback, the labeler needs to satisfy stronger conditions with respect to abstention feedback beyond Condition 2.

5.3.1 Lower Bounds

Theorem 5.1. *There is a universal constant $\delta_0 \in (0,1)$ and a labeler L satisfying Conditions 1 and 2, such that for any active learning algorithm $\mathcal{A}$, there is a $\theta^* \in [0,1]$, such that for small* *enough* ε, $\Lambda(\varepsilon, \delta_0, \mathcal{A}, L, \theta^*) \geq \Omega\left(\frac{1}{f(\varepsilon)}\varepsilon^{-2\beta}\right)$.

Theorem 5.1 establishes an information-theoretic query complexity lower bound for active learning with abstention: no algorithm can achieve an accuracy less than ε with less than $\Omega(\frac{1}{f(\varepsilon)}\varepsilon^{-2\beta})$ queries. As a comparison, [CN07] studies learning thresholds with only noisy responses, and gives a lower bound of $\Omega(\varepsilon^{-2\beta})$, which can be seen as a special case of our result.

The proof of Theorem 5.1 is similar to the one in [CN08]. We use the following formulation of Le Cam's method [Tsy08]:

Lemma 5.2. *Let Θ be a class of parameters, and $\{P_\theta : \theta \in \Theta\}$ be a class of probability distributions indexed by Θ over some sample space $\mathcal{X}$. Let $d : \Theta \times \Theta \to \mathbb{R}$ be a semi-metric. If there exist $\theta_0, \theta_1 \in \Theta$, such that $d_{KL}(P_{\theta_0}, P_{\theta_1}) \leq \alpha$ and $d(\theta_0, \theta_1) \geq 2s > 0$, then for any algorithm $\hat{\theta}$ that given a sample X outputs $\hat{\theta}(X)$, an estimation of θ, the following inequality holds:*

$$\begin{aligned} & \sup_{\theta \in \Theta} \underset{X \sim P_\theta}{P_\theta}\left(d(\theta, \hat{\theta}(X)) \geq s\right) \\ \geq \quad & \max\left\{\frac{e^{-\alpha}}{4}, \frac{1-\sqrt{\alpha/2}}{2}\right\} \end{aligned}$$

We need the following lemma in the proof of lower bounds.

Lemma 5.3. *If P, Q are distributions of two Bernoulli random variables with parameter p, q respectively and $\frac{1}{4} < p, q < \frac{1}{2}$, then $d_{KL}(P, Q) \leq 8(p-q)^2$.*

Proof.

$$\begin{aligned} d_{\mathrm{KL}}(P,Q) &= \int_q^p \left(\frac{p}{x} - \frac{1-p}{1-x}\right) dx \\ &= \int_q^p \frac{p-x}{x(1-x)} dx \\ &\leq 16 \int_q^p p - x\, dx \\ &= 8(p-q)^2 \end{aligned}$$

The inequality in line 3 follows from the fact that $x(1-x) > \frac{1}{16}$ when $\frac{1}{4} < x < \frac{1}{2}$. □

Proof of the Theorem 5.1. We take Θ be $[0,1]$, and $d(\theta_1,\theta_2) = |\theta_1 - \theta_2|$ in Lemma 5.2. We consider two thresholds $\theta_0 = 0$ and $\theta_1 = t$ where $t \in [0,1]$ is to be chosen later. Next, we will define two distributions P_0 and P_1 corresponding to P_{θ_0} and P_{θ_1} in Lemma 5.2 respectively.

For $\theta_0 = 0$, we define the distribution of labeler's response as follows:

$$P_0(Y = \perp |x) = \begin{cases} 1 - f(t) - \max\{f(x-t), f(x) - f(t)\} & x > t \\ 1 - f(t) & x \leq t \end{cases}$$

$$P_0(Y = 0|x, Y \neq \perp) = \frac{1}{2}(1 - Cx^{\beta})$$

For $\theta_1 = t$, we define the distribution of labeler's response as follows:

$$\begin{array}{rcl} P_1(Y = \perp |x) & = & P_0(Y = \perp |x) \\ & = & \begin{cases} 1 - f(t) - \max\{f(x-t), f(x) - f(t)\} & x > t \\ 1 - f(t) & x \leq t \end{cases} \end{array}$$

$$P_1(Y = 0|x, Y \neq \perp) = \begin{cases} \frac{1}{2}(1 - Cx^{\beta}) & x > t \\ \frac{1}{2}(1 + C(t-x)^{\beta}) & x \leq t \end{cases}$$

It can be checked these two distributions comply with Conditions 1 and 2.

Next, we consider P_0^n and P_1^n, the distribution of n samples $\left\{(X_i, Y_i)\right\}_{i=1}^n$ where Y_i is drawn with conditional probability P_0 and P_1 respectively, and X_i is drawn by the active learning algorithm.

$$\begin{array}{rcl} d_{\mathrm{KL}}\left(P_1^n, P_0^n\right) & = & \mathbb{E}_1\left(\log \frac{P_1^n\left(\left\{(X_i, Y_i)\right\}_{i=1}^n\right)}{P_0^n\left(\left\{(X_i, Y_i)\right\}_{i=1}^n\right)}\right) \\ & = & \mathbb{E}_{P_1}\left(\log \frac{\Pi_{i=1}^n P_1\left(Y_i|X_i\right)}{\Pi_{i=1}^n P_0\left(Y_i|X_i\right)}\middle| X_1, \ldots, X_n\right) \\ & \leq & n \max_{x \in [0,1]} \mathbb{E}_1\left(\log \frac{P_1\left(Y|x\right)}{P_0\left(Y|x\right)}\middle| x\right) \end{array}$$

where the second equality follows from the fact that the active learner will draw X_i based solely on the knowledge of $\left\{(X_j, Y_j)\right\}_{j=1}^{i-1}$, and hence $P_0\left(X_i|X_1, Y_1, X_2, Y_2, \ldots, X_{i-1}, Y_{i-1}\right) = P_1\left(X_i|X_1, Y_1, X_2, Y_2, \ldots, X_{i-1}, Y_{i-1}\right)$.

$$\begin{aligned}
& \mathbb{E}_{P_1}\left(\log \frac{P_1(Y|x)}{P_0(Y|x)}\bigg| x\right) \\
= \quad & P_1(Y=\perp|x)\log\frac{P_1(Y=\perp|x)}{P_0(Y=\perp|x)} + P_1(Y=1|x)\log\frac{P_1(Y=1|x)}{P_0(Y=1|x)} \\
& +P_1(Y=0|x)\log\frac{P_1(Y=0|x)}{P_0(Y=0|x)} \\
= \quad & 0+P_0(Y\neq\perp|x)d_{\mathrm{KL}}\left(P_1(Y|x,Y\neq\perp),P_0(Y|x,Y\neq\perp)\right) \\
\leq \quad & f(t)d_{\mathrm{KL}}\left(P_1(Y|x,Y\neq\perp),P_0(Y|x,Y\neq\perp)\right)
\end{aligned}$$

When $x\geq t$, $d_{\mathrm{KL}}\left(P_1(Y|x,Y\neq\perp),P_0(Y|x,Y\neq\perp)\right)$=0. When $x<t$, we apply Lemma 5.3 and have

$$\begin{aligned}
& d_{\mathrm{KL}}\left(P_1(Y|x,Y\neq\perp),P_0(Y|x,Y\neq\perp)\right) \\
\leq \quad & 8\left(\left(\tfrac{1}{2}(1+C(t-x)^{\beta})\right)-\left(\tfrac{1}{2}(1-Cx^{\beta})\right)\right)^2 \\
\leq \quad & 8C^2t^{2\beta}
\end{aligned}$$

In either case, we have $d_{\mathrm{KL}}\left(P_1^n,P_0^n\right)\leq 8C^2nf(t)t^{2\beta}$.

If $n\leq\frac{1}{4C^2f(t)}t^{-2\beta}$, then $d_{\mathrm{KL}}\left(P_1^n,P_0^n\right)\leq 16$. By Lemma 5.2, for any active learning algorithm, there is a $\theta\in[0,1]$, such that $P_\theta^n\left(\left|\Psi(X^n)-\theta\right|>t/2\right)>e^{-16}/4$. This concludes the proof.

□

5.3.2 Algorithm and Analysis

Next, we propose an algorithm (Algorithm 6) that works under Conditions 1 and 2. We show that this algorithm achieves an nearly optimal query complexity up to logarithmic factors and constants.

Algorithm 6 is motivated by the algorithm discussed in [CN08] which only deals with noisy labelers. It consists of two procedures: MWU and LearnThresholds. The MWU procedure is an iterative method. In each iteration, it first selects a sample to query the labeler, and then increases the weight of hypotheses that correctly label this sample and decrease the weight of those that make a mistake. The sampling strategy is generalized binary search: the algorithm selects the sample x that such that nearly half of the hypotheses assign x a label 0 and nearly half assign it label 1. If the labeler abstains from labeling, then the algorithm repeatedly queries the sample. Note that MWU only queries samples on a discrete grid Θ instead of [0,1]. In the LearnThresholds procedure, it runs MWU on three sets of grids to ensure that θ^* is far away from at least two sets of grids so that labeler's flipping adn abstention probability on these two grids is low enough for MWU to work.

The following result is a direct corollary from [CN08].

Lemma 5.4. *Suppose the labeler satisfies Condition 2 with $f(x) \equiv 1$ (i.e., no abstention). There is an absolute constant c such that if $n \geq \frac{c}{C^2}\varepsilon^{-2\beta}\log\frac{1}{\delta\varepsilon}$ and LearnThresholds(C,β,ε,n) outputs $\hat{\theta}$, then with probability at least $1-\delta$, $|\hat{\theta}-\theta^*| \leq \varepsilon$.*

In a general setting where the labeler can abstain, we have the following upper bound on the estimation error that matches the lower bound in Theorem 5.1 up to logarithmic factors and constants.

Theorem 5.5. *Suppose the labeler satisfies Condition 2. There is an absolute constant c such that*

Algorithm 6 A repetitive querying learning algorithm with a multiplicative-weight-updating subroutine.

```
procedure MWU($\gamma, \Theta = \{\theta_1, \ldots, \theta_m\}, T$)
    $p_i \leftarrow 1/m$ for $i = 0 \ldots m-1$
    $t \leftarrow 0$
    $N \leftarrow 0$
    while $t < T$ do
        $x_N \leftarrow \arg\min_i \left|\sum_{j=0}^{i} p_j - 1/2\right|$
        repeat
            Query $x_t$ and receive $y_t$
            $t \leftarrow t + 1$
        until $y_t \neq \perp$ or $t > T$
        for $i = 1, 2, \ldots, m$ do
            $p_i \leftarrow \begin{cases} p_i * (1 + 2\gamma) & \text{if } \mathbb{1}\{x_N \geq \theta_i\} = y_t \\ p_i * (1 - 2\gamma) & \text{if } \mathbb{1}\{x_N \geq \theta_i\} \neq y_t \end{cases}$
        end for
        Normalize $p$
        $N \leftarrow N + 1$
    end while
    Output: $\theta_{\text{opt}}$ where $\text{opt} = \arg\max_i p_i$
end procedure
procedure LEARNTHRESHOLDS($C, \beta, n, \varepsilon$)
    $\gamma \leftarrow C(6\varepsilon)^{\beta}$
    for $i = 0, 1, 2$ do
        $\Theta_i \leftarrow \{0 + \frac{i}{3}\varepsilon, \varepsilon + \frac{i}{3}\varepsilon, 2\varepsilon + \frac{i}{3}\varepsilon, 3\varepsilon + \frac{i}{3}\varepsilon, \ldots, \left\lfloor \frac{1}{\varepsilon} \right\rfloor \varepsilon + \frac{i}{3}\varepsilon\}$
        $\theta_i \leftarrow$ MWU($\gamma, \Theta_i, n/3$)
    end for
    for $i, j = 0, 1, 2$ do
        if $i \neq j$ and $|\theta_i - \theta_j| < \varepsilon/3$ then
            Output: $(\theta_i + \theta_j)/2$
        end if
    end for
end procedure
```

if $n \geq \frac{c}{f(\varepsilon/6)C^2}\varepsilon^{-2\beta}\log^2\frac{1}{\delta\varepsilon}$ *and* $\hat{\theta}$ *is the output of LearnThresholds(*C,β,ε,n*), then with probability at least* $1-\delta$*,* $|\hat{\theta}-\theta^*| \leq \varepsilon$.

Proof. It is easy to see there are at least 2 sets of grids (without loss of generality, let the 2 sets of grids be Θ_1 and Θ_2) that $\theta-\theta^* > \frac{\varepsilon}{6}$ for any $\theta \in \Theta_1 \cup \Theta_2$. On these two sets of grids, each query in line 8 will return a 0/1 label with probability at least $f(\frac{\varepsilon}{6})$. By the union bound, we will have with probability at least $1-\delta$, $N \geq T/(f(\frac{\varepsilon}{6})\log\frac{\delta}{2T})$ in the MWU procedure for Θ_1 and Θ_2 .

Therefore, if we set the label budget

$$n = \frac{3c}{C^2 f(\frac{\varepsilon}{6})}\left(\frac{1}{\varepsilon}\right)^{2\beta}\log\frac{1}{\varepsilon\delta}\log\left(\frac{1}{C^2}\left(\frac{1}{\varepsilon}\right)^{2\beta}\log\frac{1}{\varepsilon\delta}\right),$$

for Θ_1 and Θ_2, the number of non-abstaining responses $N \geq \frac{c}{C^2}\left(\frac{1}{\varepsilon}\right)^{2\beta}\log\frac{1}{\varepsilon\delta}$ with probability at least $1-\delta/2$. Consequently by Lemma 5.4 we will have $|\theta_1-\theta^*| \leq \varepsilon$ and $|\theta_2-\theta^*| \leq \varepsilon$ with probability at least $1-\delta$. Thus, LearnThresholds in Algorithm 6 will output a $\hat{\theta}$ such that $|\hat{\theta}-\theta^*| \leq \varepsilon$ with probability at least $1-\delta$. This concludes the proof. □

Algorithm 6 achieves a nearly optimal query complexity of $\tilde{\Theta}(\frac{1}{f(\varepsilon)}\varepsilon^{-2\beta})$ by simply ignoring abstention feedback. Therefore, if we would like to improve the query complexity of the algorithm by making use of abstention feedback, the labeler needs to satisfy stronger conditions with respect to abstention feedback beyond Conditions 1 and 2.

5.4 Active Learning with Monotonic Abstention Rates

In this section, we consider active learning under Conditions 1, 2, and 3 where the abstention rate is, roughly speaking, strictly monotonic. We provide an active learning algorithm

(Algorithm 7) that exploits the abstention feedback under these conditions. We prove that this algorithm is statistically consistent under the very mild Condition 1. It achieves the same query complexity as that for Algorithm 6 under Conditions 1 and 2. Under Conditions 1, 2, and 3, we show that it achieves substantially better query complexity. More importantly, unlike Algorithm 6, Algorithm 7 is completely adaptive to parameters of the labeler (C, β, f).

5.4.1 Algorithm

The proposed algorithm is a binary search style algorithm shown as Algorithm 7. (For the sake of simplicity, we assume $\log \frac{1}{2\varepsilon}$ is an integer.) Algorithm 7 takes a desired precision ε and confidence level δ as its input, and returns an estimation $\hat{\theta}$ of the decision boundary θ^*. The algorithm maintains an interval $[L_k, R_k]$ in which θ^* is believed to lie, and shrinks this interval iteratively. To find the subinterval that contains θ^*, Algorithm 7 relies on two auxiliary functions (marked in Procedure 8) to conduct adaptive sequential hypothesis tests regarding subintervals of interval $[L_k, R_k]$.

Suppose $\theta^* \in [L_k, R_k]$. Algorithm 7 tries to shrink this interval to a $\frac{3}{4}$ of its length in each iteration by repetitively querying on quartiles $U_k = \frac{3L_k+R_k}{4}$, $M_k = \frac{L_k+R_k}{2}$, $V_k = \frac{L_k+3R_k}{4}$. To determine which specific subinterval to choose, the algorithm uses 0/1 labels and abstention responses simultaneously. Since the ground truth labels are determined by $\mathbb{1}[x > \theta^*]$, one can infer that if the number of queries that return label 0 at U_k (V_k) is statistically significantly more (less) than label 1, then θ^* should be on the right (left) side of U_k (V_k). Similarly, from Condition 1, if the number of non-abstention responses at U_k (V_k) is statistically significantly more than non-abstention responses at M_k, then θ^* should be closer to M_k than U_k (V_k).

Algorithm 7 relies on the ability to shrink the search interval via statistically comparing the numbers of obtained labels at locations U_k, M_k, V_k. As a result, a main building block of

Algorithm 7 The active learning algorithm for learning thresholds

```
Input: $\delta$, $\varepsilon$
$[L_0, R_0] \leftarrow [0, 1]$
for $k = 0, 1, 2, \ldots, \log\frac{1}{2\varepsilon} - 1$ do
    Define three quartiles: $U_k \leftarrow \frac{3L_k+R_k}{4}$, $M_k \leftarrow \frac{L_k+R_k}{2}$, $V_k \leftarrow \frac{L_k+3R_k}{4}$
    $A^{(u)}, A^{(m)}, A^{(v)}, B^{(u)}, B^{(v)} \leftarrow$ Empty Array
    for $n = 1, 2, \ldots$ do
        Query at $U_k, M_k, V_k$, and receive labels $X_n^{(u)}, X_n^{(m)}, X_n^{(v)}$
        for $w \in \{u, m, v\}$ do
            ▷ We record whether $X^{(w)} = \perp$ in $A^{(w)}$, and the 0/1 label (as -1/1) in $B^{(w)}$ if $X^{(w)} \neq \perp$
            if $X^{(w)} \neq \perp$ then
                $A^{(w)} \leftarrow A^{(w)}$.append(1) , $B^{(w)} \leftarrow B^{(w)}$.append($2\mathbb{1}\left[X^{(w)} = 1\right] - 1$)
            else
                $A^{(w)} \leftarrow A^{(w)}$.append(0)
            end if
        end for
        ▷ Check if the differences of abstention responses are statistically significant
        if CheckSignificant-Var($\left\{A_i^{(u)} - A_i^{(m)}\right\}_{i=1}^{n}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}$) then
            $[L_{k+1}, R_{k+1}] \leftarrow [U_k, R_k]$; break
        else if CheckSignificant-Var($\left\{A_i^{(v)} - A_i^{(m)}\right\}_{i=1}^{n}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}$) then
            $[L_{k+1}, R_{k+1}] \leftarrow [L_k, V_k]$; break
        end if
        ▷ Check if the differences between 0 and 1 labels are statistically significant
        if CheckSignificant($\left\{-B_i^{(u)}\right\}_{i=1}^{B^{(u)}.\text{length}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}$) then
            $[L_{k+1}, R_{k+1}] \leftarrow [U_k, R_k]$; break
        else if CheckSignificant($\left\{B_i^{(v)}\right\}_{i=1}^{B^{(v)}.\text{length}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}$) then
            $[L_{k+1}, R_{k+1}] \leftarrow [L_k, V_k]$; break
        end if
    end for
end for
Output: $\hat{\theta} = \left(L_{\log\frac{1}{2\varepsilon}} + R_{\log\frac{1}{2\varepsilon}}\right)/2$
```

Procedure 8 Adaptive sequential testing

```
▷ $D_0, D_1$ are absolute constants defined in Proposition 5.6 and Proposition 5.7
▷ $\{X_i\}$ are i.i.d. random variables bounded by 1. $\delta$ is the confidence level. Detect if $\mathbb{E}X > 0$
function CheckSignificant($\{X_i\}_{i=1}^{n}, \delta$)
    $p(n,\delta) \leftarrow D_0 \left(1 + \ln\frac{1}{\delta} + \sqrt{4n\left([\ln\ln]_+ 4n + \ln\frac{1}{\delta}\right)}\right)$
    Return $\sum_{i=1}^{n} X_i \geq p(n,\delta)$
end function
function CheckSignificant-Var($\{X_i\}_{i=1}^{n}, \delta$)
    Calculate the empirical variance $\mathrm{Var} = \frac{n}{n-1}\left(\sum_{i=1}^{n} {X_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right)$
    $q(n, \mathrm{Var}, \delta) \leftarrow D_1 \left(1 + \ln\frac{1}{\delta} + \sqrt{\left(\mathrm{Var} + \ln\frac{1}{\delta} + 1\right)\left([\ln\ln]_+ \left(\mathrm{Var} + \ln\frac{1}{\delta} + 1\right) + \ln\frac{1}{\delta}\right)}\right)$
    Return $n \geq \ln\frac{1}{\delta}$ AND $\sum_{i=1}^{n} X_i \geq q(n, \mathrm{Var}, \delta)$
end function
```

Algorithm 7 is to test whether i.i.d. bounded random variables Y_i are greater in expectation than i.i.d. bounded random variables Z_i with statistical significance. In Procedure 8, we have two test functions CheckSignificant and CheckSignificant-Var that take i.i.d. random variables $\{X_i = Y_i - Z_i\}$ ($|X_i| \leq 1$) and confidence level δ as their input, and output whether it is statistically significant to conclude $\mathbb{E}X_i > 0$.

CheckSignificant is based on the following uniform concentration result regarding the empirical mean:

Proposition 5.6. *Suppose $X_1, X_2, \ldots$ are a sequence of i.i.d. random variables with $X_1 \in [-2,2]$, $\mathbb{E}X_1 = 0$. Take any $0 < \delta < 1$. Then there is an absolute constant D_0 such that with probability at least $1-\delta$, for all $n > 0$ simultaneously,*

$$\left|\sum_{i=1}^{n} X_i\right| \leq D_0 \left(1 + \ln\frac{1}{\delta} + \sqrt{4n\left([\ln\ln]_+ 4n + \ln\frac{1}{\delta}\right)}\right)$$

In Algorithm 7, we use CheckSignificant to detect whether the expected number of queries that return label 0 at location U_k (V_k) is more/less than the expected number of label 1 with a statistical significance.

CheckSignificant-Var is based on the following uniform concentration result which further utilizes the empirical variance $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n X_i^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2\right)$:

Proposition 5.7. *There is an absolute constant* D_1 *such that with probability at least* $1-\delta$, *for all* $n \geq \ln\frac{1}{\delta}$ *simultaneously,*

$$\left|\sum_{i=1}^n X_i\right| \leq D_1\left(1+\ln\frac{1}{\delta}+\sqrt{\left(1+\ln\frac{1}{\delta}+V_n\right)\left([\ln\ln]_+(1+\ln\frac{1}{\delta}+V_n)+\ln\frac{1}{\delta}\right)}\right)$$

The use of variance results in a tighter bound when $\mathrm{Var}(X_i)$ is small.

In Algorithm 7, we use CheckSignificant-Var to detect the statistical significance of the relative order of the number of queries that return non-abstention responses at U_k (V_k) compared to the number of non-abstention responses at M_k. This results in a better query complexity than using CheckSignificant under Condition 3, since the variance of the number of abstention responses approaches 0 when the interval $[L_k, R_k]$ zooms in on θ^*.[1]

5.4.2 Analysis

In this subsection, we use $\log x = \log_{\frac{4}{3}} x$ for convenience since the proposed algorithm shrinks the search interval by a factor of $\frac{3}{4}$ at each time.

[1] We do not apply CheckSignificant-Var to 0/1 labels, because unlike the difference between the numbers of abstention responses at U_k (V_k) and M_k, the variance of the difference between the numbers of 0 and 1 labels stays above a positive constant.

Properties of adaptive sequential testing in Procedure 8

Lemma 5.8. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i \leq 0$, $|X_i| \leq 1$. Let $\delta > 0$. Then for CheckSignificant$\left(\{X_i\}_{i=1}^{n}, \delta\right)$ in Procedure 8, with probability at least $1-\delta$, it returns false for all $n \in \mathbb{N}$ simultaneously.*

Proof. This is immediate by applying Proposition 5.6 to $X_i - \mathbb{E}X_i$. □

Lemma 5.9. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i > \varepsilon > 0$, $|X_i| \leq 1$. Let $\delta \in [0, \frac{1}{3}]$, $N \geq \frac{\xi}{\varepsilon^2} \ln \frac{1}{\delta} [\ln \ln]_+ \frac{1}{\varepsilon}$ (ξ is an absolute constant specified in the proof). Then with probability at least $1-\delta$, CheckSignificant$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ in Procedure 8 returns true.*

Proof. Let $S_N = \sum_{i=1}^{N} X_i$. CheckSignificant$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ returns false if and only if $S_N \leq D_0 \left(1 + \ln \frac{1}{\delta} + \sqrt{N\left([\ln \ln]_+ N + \ln \frac{1}{\delta}\right)}\right)$.

$$\begin{aligned}
&\mathbb{P}\left(S_N \leq D_0 \left(1 + \ln \frac{1}{\delta} + \sqrt{N\left([\ln \ln]_+ N + \ln \frac{1}{\delta}\right)}\right)\right) \\
\leq &\mathbb{P}\left(S_N \leq D_0 \left(1 + \ln \frac{1}{\delta} + \sqrt{N[\ln \ln]_+ N} + \sqrt{N \ln \frac{1}{\delta}}\right)\right) \\
\leq &\mathbb{P}\left(S_N - N\mathbb{E}X_i \leq D_0 \left(1 + \ln \frac{1}{\delta} + \sqrt{N[\ln \ln]_+ N} + \sqrt{N \ln \frac{1}{\delta}}\right) - N\varepsilon\right)
\end{aligned}$$

Suppose $N = \frac{c\xi}{\varepsilon^2} \ln \frac{1}{\delta} [\ln \ln]_+ \frac{1}{\varepsilon}$ for constant $c \geq 1$ and ξ. ξ is set to be sufficiently large, such that (1) $\xi \geq 4D_0^2$; (2) $\frac{2D_0}{\sqrt{\xi}} + D_0 \left(3 + \sqrt{[\ln \ln]_+ \xi}\right) + D_0 - \sqrt{\xi}/2 \leq -\sqrt{\frac{1}{2}}$; and (3) $f(x) = D_0 \sqrt{[\ln \ln]_+ x} - \sqrt{x}/2$ is decreasing when $x > \xi$. Here (2) is satisfiable since $\frac{D_0}{\sqrt{\xi}} + D_0 \sqrt{[\ln \ln]_+ \xi} -$

$\sqrt{\xi}/2 \to -\infty$ as $\xi \to \infty$, (3) is satisfiable since $f'(x) \to -\infty$ as $x \to \infty$. (2) and (3) together implies $\frac{2D_0}{\sqrt{\xi}} + D_0\left(3+\sqrt{[\ln\ln]_+ c\xi}\right) + D_0 - \sqrt{c\xi}/2 \le -\sqrt{\frac{1}{2}}$.

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\left(D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right)-N\varepsilon\right)\\
=&\sqrt{\ln\frac{1}{\delta}}\left(\frac{D_0\varepsilon(1+\ln\frac{1}{\delta})}{\sqrt{c\xi[\ln\ln]_+\frac{1}{\varepsilon}\ln\frac{1}{\delta}}}+D_0\sqrt{\frac{[\ln\ln]_+\left(\frac{c\xi}{\varepsilon^2}\ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\varepsilon}\right)}{\ln\frac{1}{\delta}}}+D_0-\sqrt{c\xi[\ln\ln]_+\frac{1}{\varepsilon}}\right)
\end{aligned}
$$

Since $[\ln\ln]_+\frac{1}{\varepsilon}, c, \ln\frac{1}{\delta} \ge 1$ and $\varepsilon < 1$, we have $\frac{D_0\varepsilon(1+\ln\frac{1}{\delta})}{\sqrt{c\xi[\ln\ln]_+\frac{1}{\varepsilon}\ln\frac{1}{\delta}}} \le \frac{2D_0}{\sqrt{\xi}}$.

Since $[\ln\ln]_+ x \ge 1$ if $x \ge 1$, we have $[\ln\ln]_+\frac{1}{\varepsilon} \le \frac{1}{\varepsilon}$, and thus

$$
\begin{aligned}
\sqrt{[\ln\ln]_+\left(\frac{c\xi}{\varepsilon^2}\ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\varepsilon}\right)} &= \sqrt{\ln\left[\max\left\{e, 2\ln\frac{1}{\varepsilon}+\ln c\xi+\ln\ln\frac{1}{\delta}+\ln[\ln\ln]_+\frac{1}{\varepsilon}\right\}\right]}\\
&\le \sqrt{\ln\left[\max\left\{e, 3\ln\frac{1}{\varepsilon}+\ln c\xi+[\ln\ln]_+\frac{1}{\delta}\right\}\right]}\\
&\overset{(a)}{\le} \sqrt{\ln\left[\max\left\{e, 9\ln\frac{1}{\varepsilon}\ln c\xi[\ln\ln]_+\frac{1}{\delta}\right\}\right]}\\
&\le \sqrt{3+[\ln\ln]_+\frac{1}{\varepsilon}+[\ln\ln]_+c\xi+\ln[\ln\ln]_+\frac{1}{\delta}}\\
&\overset{(b)}{\le} \sqrt{3}+\sqrt{[\ln\ln]_+c\xi}+\sqrt{[\ln\ln]_+\frac{1}{\varepsilon}}+\sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}
\end{aligned}
$$

where (a) follows by $a+b+c \le 3abc$ if $a,b,c \ge 1$, and (b) follows by $\sqrt{\sum_i x_i} \le \sum_i \sqrt{x_i}$ if $x_i \ge 0$.

Thus, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left(D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right)-N\varepsilon\right)\\
\leq&\sqrt{\ln\frac{1}{\delta}}\left(\frac{2D_0}{\sqrt{\xi}}+D_0\frac{\sqrt{3}+\sqrt{[\ln\ln]_+c\xi}+\sqrt{[\ln\ln]_+\frac{1}{\varepsilon}}+\sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}}{\sqrt{\ln\frac{1}{\delta}}}+D_0-\sqrt{c\xi[\ln\ln]_+\frac{1}{\varepsilon}}\right)\\
\overset{(c)}{\leq}&\sqrt{\ln\frac{1}{\delta}}\left(\frac{2D_0}{\sqrt{\xi}}+D_0\left(3+\sqrt{[\ln\ln]_+c\xi}\right)+D_0-\sqrt{c\xi}/2\right)\\
\overset{(d)}{\leq}&-\sqrt{\ln\frac{1}{\delta}/2}
\end{aligned}$$

(c) follows by $\sqrt{\ln\frac{1}{\delta}}\geq\max\left\{1,\sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}\right\}$, $D_0\geq 1$, and $\sqrt{[\ln\ln]_+\frac{1}{\varepsilon}}(\frac{D_0}{\sqrt{\ln\frac{1}{\delta}}}-\sqrt{c\xi})\leq D_0-\sqrt{c\xi}\leq-\sqrt{c\xi}/2$ if $c\xi\geq 4D_0^2$. (d) follows by our choose of ξ.

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(S_N-N\mathbb{E}X_i\leq D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right)-N\varepsilon\right)\\
\leq&\mathbb{P}\left(S_N-N\mathbb{E}X_i\leq-\sqrt{N\ln\frac{1}{\delta}/2}\right)
\end{aligned}$$

which is at most δ by Hoeffding Bound. □

Lemma 5.10. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i\leq 0$, $|X_i|\leq 1$. Let $\delta>0$. Then with probability at least $1-\delta$, for all n simultaneously CheckSignificant-Var$\big(\{X_i\}_{i=1}^{n},\delta\big)$ in Procedure 8 returns false.*

Proof. Define $Y_i=X_i-\mathbb{E}X_i$. It is easy to check $\frac{n}{n-1}(\sum_{i=1}^{n}Y_i^2-\frac{1}{n}(\sum_{i=1}^{n}Y_i)^2)=\frac{n}{n-1}(\sum_{i=1}^{n}X_i^2-$

$\frac{1}{n}(\sum_{i=1}^{n} X_i)^2)$. The result is immediate from Proposition 5.7. □

Lemma 5.11. *Suppose* $\{X_i\}_{i=1}^{\infty}$ *is a sequence of i.i.d. random variables such that* $\mathbb{E}X_i > \tau\varepsilon$, $|X_i| \leq 1$, $Var(X_i) \leq 2\varepsilon$ *where* $0 < \varepsilon \leq 1$, $\tau > 0$. *Let* $\delta < 1$, $N = \frac{\xi}{\tau\varepsilon}\ln\frac{2}{\delta}$ *(*ξ *is a constant specified in the proof). Then with probability at least* $1-\delta$, *CheckSignificant-Var*$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ *in Procedure 8 returns true.*

Proof. Let $Y_i = X_i - \mathbb{E}X_i$, η be the constant η in Lemma B.9. Set $\xi = \max(\eta, \frac{16}{\tau} + \frac{8}{3})$.

CheckSignificant-Var$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ returns false if and only if $\sum_{i=1}^{N} X_i \leq q(N, \text{Var}, \delta)$.

By applying Lemma B.9 to X_i, $\frac{q(N,\text{Var},\delta)}{N} - \mathbb{E}X_i \leq -\tau\varepsilon/2$ with probability at least $1-\delta/2$.

Applying Bernstein's inequality to Y_i, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \leq -\tau\varepsilon/2\right) &\leq \exp\left(-\frac{N(-\tau\varepsilon)^2/4}{4\varepsilon + 2\tau\varepsilon/3}\right)\\ &= \exp\left(-\frac{\xi\ln\frac{2}{\delta}}{16/\tau + 8/3}\right)\\ &\leq \delta/2\end{aligned}$$

Thus, by a union bound,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{N} X_i \le q(N,\mathrm{Var},\delta)\right)\\
\le&\mathbb{P}\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \ge -\tau\varepsilon/2\right)\\
&+\mathbb{P}\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \le -\tau\varepsilon/2 \text{ and } \frac{1}{N}\sum_{i=1}^{N} X_i \le \frac{q(N,\mathrm{Var},\delta)}{N}\right)\\
\le&\delta/2+\mathbb{P}\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \le -\tau\varepsilon/2 \text{ and } \frac{1}{N}\sum_{i=1}^{N} Y_i \le \frac{q(n,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i\right)\\
\le&\delta/2+\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \le -\tau\varepsilon/2\right)\\
\le&\delta
\end{aligned}
$$

□

Consistency

For Algorithm 7 to be statistically consistent, we only need Condition 1.

Theorem 5.12. *Let θ^* be the ground truth. If the labeler L satisfies Condition 1 and Algorithm 7 stops to output $\hat{\theta}$, then $\left|\theta^* - \hat{\theta}\right| \le \varepsilon$ with probability at least $1-\frac{\delta}{2}$.*

Proof. Since $\hat{\theta} = \left(L_{\log\frac{1}{2\varepsilon}} + R_{\log\frac{1}{2\varepsilon}}\right)/2$ and $R_{\log\frac{1}{2\varepsilon}} - L_{\log\frac{1}{2\varepsilon}} = 2\varepsilon$, $\left|\hat{\theta} - \theta^*\right| > \varepsilon$ is equivalent to $\theta^* \notin [L_{\log\frac{1}{2\varepsilon}}, R_{\log\frac{1}{2\varepsilon}}]$. We have

$$
\begin{aligned}
\mathbb{P}\left(\left|\hat{\theta}-\theta^*\right| > \varepsilon\right) &= \mathbb{P}\left(\theta^* \notin [L_{\log\frac{1}{2\varepsilon}}, R_{\log\frac{1}{2\varepsilon}}]\right)\\
&= \mathbb{P}\left(\exists k : \theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right)\\
&\le \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} \mathbb{P}\left(\theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right)
\end{aligned}
$$

For any $k = 0, \ldots, \log \frac{1}{2\varepsilon} - 1$, define $\mathbb{Q}_k = \left\{ (p,q) : p,q \in \mathbb{Q} \cap [0,1] \text{ and } q - p = \left(\frac{3}{4}\right)^k \right\}$ where $\mathbb{Q}$ is the set of rational numbers. Note that $L_k, R_k \in \mathbb{Q}_k$, and $\mathbb{Q}$ is countable. So we have

$$\begin{aligned}
&\mathbb{P}\left(\theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right) \\
= &\sum_{(p,q) \in \mathbb{Q}_k : p \leq \theta^* \leq q} \mathbb{P}\left(L_k = p, R_k = q \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right) \\
= &\sum_{(p,q) \in \mathbb{Q}_k : p \leq \theta^* \leq q} \mathbb{P}\left(\theta^* \notin [L_{k+1}, R_{k+1}] | L_k = p, R_k = q\right) \mathbb{P}\left(L_k = p, R_k = q\right)
\end{aligned}$$

Define event $E_{k,p,q}$ to be the event $L_k = p, R_k = q$. To show $\mathbb{P}\left(\left|\hat{\theta} - \theta^*\right| > \varepsilon\right) \leq \frac{\delta}{2}$, it suffices to show $\mathbb{P}\left(\theta^* \notin [L_{k+1}, R_{k+1}] | E_{k,p,q}\right) \leq \frac{\delta}{2 \log \frac{1}{2\varepsilon}}$ for any $k = 0, \ldots, \log \frac{1}{2\varepsilon} - 1$, $(p,q) \in \mathbb{Q}_k$ and $p \leq \theta^* \leq q$.

Conditioning on event $E_{k,p,q}$, event $\theta^* \notin [L_{k+1}, R_{k+1}]$ happens only if some calls of CheckSignificant and CheckSignificant-Var between Line 16 and 27 of Algorithm 7 return true incorrectly. In other words, at least one of following events happens for some n:

- $O^{(1)}_{k,p,q}$: $\theta^* \in [L_k, U_k]$ and CheckSignificant-Var($\left\{A^{(u)}_i - A^{(m)}_i\right\}_{i=1}^n, \frac{\delta}{4 \log \frac{1}{2\varepsilon}}$) returns true;
- $O^{(2)}_{k,p,q}$: $\theta^* \in [V_k, R_k]$ and CheckSignificant-Var($\left\{A^{(v)}_i - A^{(m)}_i\right\}_{i=1}^n, \frac{\delta}{4 \log \frac{1}{2\varepsilon}}$) returns true;
- $O^{(3)}_{k,p,q}$: $\theta^* \in [L_k, U_k]$ and CheckSignificant($\left\{-B^{(u)}_i\right\}_{i=1}^n, \frac{\delta}{4 \log \frac{1}{2\varepsilon}}$) returns true;
- $O^{(4)}_{k,p,q}$: $\theta^* \in [V_k, R_k]$ and CheckSignificant($\left\{B^{(v)}_i\right\}_{i=1}^n, \frac{\delta}{4 \log \frac{1}{2\varepsilon}}$) returns true;

Note that since $[U_k, V_k] \subset [L_{k+1}, R_{k+1}]$ for any k by our construction, if $\theta^* \in [U_k, V_k]$ then $\theta^* \in [L_{k+1}, R_{k+1}]$. Besides, event $\theta^* \in [L_k, U_k]$ and event $\theta^* \in [V_k, R_k]$ are mutually exclusive.

Conditioning on event $E_{k,p,q}$, suppose for now $\theta^* \in [L_k, U_k]$.

$$\begin{aligned}
&\mathbb{P}\left(O^{(1)}_{k,p,q} \mid E_{k,p,q}\right) \\
=&\mathbb{P}\left(\exists n : \text{CheckSignificant-Var}(\left\{D_i^{(u,m)}\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\varepsilon}}) \text{ returns true} \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right)
\end{aligned}$$

On event $\theta^* \in [L_k, U_k]$ and $E_{k,p,q}$, the sequences $\left\{A_i^{(u)}\right\}$ and $\left\{A_i^{(m)}\right\}$ are i.i.d., and $\mathbb{E}\left[A_i^{(u)} - A_i^{(m)} \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right] \leq 0$. By Lemma 5.10, the probability above is at most $\frac{\delta}{4\log\frac{1}{2\varepsilon}}$.

Likewise,

$$\begin{aligned}
&\mathbb{P}\left(O^{(3)}_{k,p,q} \mid E_{k,p,q}\right) \\
=&\mathbb{P}\left(\exists n : \text{CheckSignificant}(\left\{-B_i^{(u)}\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\varepsilon}}) \text{ returns true} \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right)
\end{aligned}$$

On event $\theta^* \in [L_k, U_k]$ and $E_{k,p,q}$, the sequence $\left\{B_i^{(u)}\right\}$ is i.i.d., and $\mathbb{E}[-B_i^{(u)} \mid \theta^* \in [L_k, U_k], E_{k,p,q}] \leq 0$. By Lemma 5.8, the probability above is at most $\frac{\delta}{4\log\frac{1}{2\varepsilon}}$.

Thus, $\mathbb{P}\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$ when $\theta^* \in [L_k, U_k]$. Similarly, when $\theta^* \in [V_k, R_k]$, we can show $\mathbb{P}\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \leq \mathbb{P}\left(O^{(2)}_{k,p,q} \mid E_{k,p,q}\right) + \mathbb{P}\left(O^{(4)}_{k,p,q} \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$.

Therefore, $\mathbb{P}\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$, and thus $\mathbb{P}\left(\left|\hat{\theta} - \theta^*\right| > \varepsilon\right) \leq \delta/2$. □

Query Complexity Upper Bounds

Under additional Conditions 2 and 3, we can derive upper bounds of the query complexity for our algorithm. (Recall f and β are defined in Conditions 2 and 3.)

Theorem 5.13. *Let θ^* be the ground truth, and $\hat{\theta}$ be the output of Algorithm 7. Under Conditions 1 and 2, with probability at least $1-\delta$, Algorithm 7 makes at most $\tilde{O}\left(\frac{1}{f(\frac{\varepsilon}{2})}\varepsilon^{-2\beta}\right)$ queries.*

Proof. Define T_k to be the number of iterations of the loop at Line 6, $T = \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} T_k$. For any numbers $m_1, m_2, \ldots, m_{\log\frac{1}{2\varepsilon}-1}$, we have:

$$\begin{aligned}
\mathbb{P}(T \geq m) &\leq \mathbb{P}\left(\left|\hat{\theta}-\theta^*\right| > \varepsilon\right) + \mathbb{P}\left(\left|\hat{\theta}-\theta^*\right| < \varepsilon \text{ and } T \geq \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} m_k\right) \\
&\leq \frac{\delta}{2} + \mathbb{P}\left(T \geq \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} m_k \text{ and } \left|\hat{\theta}-\theta^*\right| < \varepsilon\right) \qquad (5.1) \\
&\leq \frac{\delta}{2} + \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} \mathbb{P}\left(T_k \geq m_k \text{ and } \left|\hat{\theta}-\theta^*\right| < \varepsilon\right) \\
&\leq \frac{\delta}{2} + \sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} \mathbb{P}\left(T_k \geq m_k \text{ and } \theta^* \in [L_k, R_k]\right)
\end{aligned}$$

The first and the third inequality follows by union bounds. The second follows by Theorem 5.12. The last follows since $\left|\hat{\theta}-\theta^*\right| < \varepsilon$ is equivalent to $\theta^* \in [L_{\log\frac{1}{2\varepsilon}}, R_{\log\frac{1}{2\varepsilon}}]$, which implies $\theta^* \in [L_k, R_k]$ for all $k = 0, \ldots, \log\frac{1}{2\varepsilon}-1$.

We define $\mathbb{Q}_k$ as in the previous proof. For all $k = 0, \ldots, \log\frac{1}{2\varepsilon}-1$,

$$
\begin{aligned}
&\mathbb{P}\left(T_k \geq m_k \text{ and } \theta^* \in [L_k, R_k]\right) \\
= &\sum_{(p,q)\in\mathbb{Q}_k : p \leq \theta^* \leq q} \mathbb{P}\left(T_k \geq m_k, L_k = p, R_k = q\right) \\
= &\sum_{(p,q)\in\mathbb{Q}_k : p \leq \theta^* \leq q} \mathbb{P}\left(T_k \geq m_k | L_k = p, R_k = q\right)\mathbb{P}\left(L_k = p, R_k = q\right)
\end{aligned}
$$

Thus, in order to prove the query complexity of Algorithm 7 is $O\left(\sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} m_k\right)$, it suffices to show that $\mathbb{P}\left(T_k \geq m_k \mid L_k = p, R_k = q\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$ for any $k = 0, \ldots, \log\frac{1}{2\varepsilon} - 1$, $(p,q) \in \mathbb{Q}_k$ and $p \leq \theta^* \leq q$.

For each k, p, q, define event $E_{k,p,q}$ to be the event $L_k = p, R_k = q$. Define $l_k = q - p = \left(\frac{3}{4}\right)^k$, N_k to be $\tilde{\Theta}\left(\frac{1}{f(l_k/4)} l_k^{-2\beta}\right)$. The logarithm factor of N_k is to be specified later. Define $S_n^{(u)}$ and $S_n^{(v)}$ to be the size of array $B^{(u)}$ and $B^{(v)}$ before Line 16 respectively.

To show $\mathbb{P}\left(T_k \geq N_k \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$, it suffices to show that on event $E_{k,p,q}$, with probability at least $1 - \frac{\delta}{2\log\frac{1}{2\varepsilon}}$, if $n = N_k$ then at least one of the two calls to CheckSignificant between Line 22 and Line 27 will return true.

On event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$ (note that on event $E_{k,p,q}$, L_k and M_k are deterministic), then $|V_k - \theta^*| \geq \frac{l_k}{4}$. We will show

$$
p_1 := \mathbb{P}\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right) \text{ returns false} \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}
$$

To prove this, we will first show that $S_{N_k}^{(v)}$, the length of the array $B^{(v)}$, is large with high probability, and then apply Lemma 5.9 to show that CheckSignificant will return true if $S_{N_k}^{(v)}$ is large.

By definition, $S_{N_k}^{(v)} = \sum_{i=1}^{N_k} A_i^{(v)}$. By Condition 2, we have $\mathbb{E}[A_i^{(v)} \mid E_{k,p,q}] = \mathbb{P}(Y \neq \perp \mid X = V_k, E_{k,p,q}) \geq f(\frac{l_k}{4})$.

On event $E_{k,p,q}$, $\left\{A_i^{(v)}\right\}$ is a sequence of i.i.d. random variables. By the multiplicative Chernoff bound, $\mathbb{P}\left(S_{N_k}^{(v)} \leq \frac{1}{2}N_kf\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \leq \exp\left(-N_kf\left(\frac{l_k}{4}\right)/8\right)$.

Now,

$$\begin{aligned} p_1 \leq & \mathbb{P}\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right) \text{ returns false}, S_{N_k}^{(v)} \geq \frac{1}{2}N_kf\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \\ & + \mathbb{P}\left(S_{N_k}^{(v)} < \frac{1}{2}N_kf\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \end{aligned}$$

By Condition 2 and $|V_k - \theta^*| \geq \frac{l_k}{4}$, $\mathbb{E}\left[B_i^{(v)} \mid E_{k,p,q}\right] \geq C\left(\frac{l_k}{4}\right)^\beta$. On event $E_{k,p,q}$, $\left\{B_i^{(v)}\right\}$ is a sequence of i.i.d. random variables. Thus, On event $E_{k,p,q}$, by Lemma 5.9, with probability at least $1 - \frac{\delta}{4\log\frac{1}{2\varepsilon}}$, CheckSignificant will return true if $\frac{1}{2}N_kf\left(\frac{l_k}{4}\right) = \Theta\left(\frac{1}{l_k^{2\beta}}\ln\frac{\ln 1/\varepsilon}{\delta}[\ln\ln]_+\frac{1}{l_k^{2\beta}}\right)$. We have already proved $\mathbb{P}\left(S_{N_k}^{(v)} \leq \frac{1}{2}N_kf\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \leq \exp\left(-N_kf\left(\frac{l_k}{4}\right)/8\right)$. By setting $N_k = \Theta\left(\frac{1}{f(l_k/4)}l_k^{-2\beta}\ln\frac{\ln 1/\varepsilon}{\delta}[\ln\ln]_+\frac{1}{l_k^{2\beta}}\right)$, we can ensure p_1 is at most $\delta/2\log\frac{1}{2\varepsilon}$.

Now we have proved on event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$, then

$$\mathbb{P}\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right) \text{ returns true} \mid E_{k,p,q}\right) \geq 1 - \frac{\delta}{2\log\frac{1}{2\varepsilon}}$$

Likewise, on event $E_{k,p,q}$, if $\theta^* \in [M_k, R_k]$, then

$$\mathbb{P}\left(\text{CheckSignificant}\left(\left\{-B_i^{(u)}\right\}_{i=1}^{S_{N_k}^{(u)}}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right) \text{ returns true} \mid E_{k,p,q}\right) \geq 1 - \frac{\delta}{2\log\frac{1}{2\varepsilon}}$$

Therefore, we have shown $\mathbb{P}\left(T_k \geq N_k \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$ for any k, p, q. By (5.1), with probability at least $1-\delta$, the number of samples queried is at most

$$\sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} O\left(\frac{1}{f(\left(\frac{3}{4}\right)^k/4)}\left(\frac{3}{4}\right)^{-2\beta k}\ln\frac{\ln 1/\varepsilon}{\delta}[\ln\ln]_+\left(\frac{3}{4}\right)^{-2k\beta}\right)$$
$$=O\left(\frac{\varepsilon^{-2\beta}}{f(\varepsilon/2)}\ln\frac{1}{\varepsilon}\left(\ln\frac{1}{\delta}+\ln\ln\frac{1}{\varepsilon}\right)[\ln\ln]_+\frac{1}{\varepsilon}\right)$$

□

Theorem 5.14. *Let θ^* be the ground truth, and $\hat{\theta}$ be the output of Algorithm 7. Under Conditions 1 and 3, with probability at least $1-\delta$, Algorithm 7 makes at most $\tilde{O}\left(\frac{1}{f(\frac{\varepsilon}{2})}\right)$ queries.*

Proof of Theorem 5.14. For each k in Algorithm 7 at Line 3, Let $l_k = R_k - L_k$. Let $N_k = \eta\frac{1}{f(l_k/4)}\ln\frac{4\log\frac{1}{2\varepsilon}}{\delta}$, where η is a constant to be specified later. As with the previous proof, it suffices to show $\mathbb{P}\left(T_k \geq N_k \mid E_{k,p,q}\right) \leq \frac{\delta}{2\log\frac{1}{2\varepsilon}}$ where event $E_{k,p,q}$ is defined to be $L_k = p, R_k = q$, T_k is the number of iterations at the loop at Line 6.

On event $E_{k,p,q}$, we will show that the loop at Line 6 will terminate after $n = N_k$ with probability at least $1 - \frac{\delta}{2\log\frac{1}{2\varepsilon}}$.

Suppose for now $\theta^* \in [M_k, R_k]$. Let $Z_i = A_i^{(u)} - A_i^{(m)}$, $\zeta = \theta^* - M_k$. Clearly, $|Z_i| \leq 1$. On event $E_{k,p,q}$, sequence $\{Z_i\}$ is i.i.d.. By Condition 3, $\mathbb{E}\left[Z_i \mid E_{k,p,q}\right] = f(\zeta + \frac{l_k}{4}) - f(\zeta) \geq cf(\zeta + \frac{l_k}{4})$ since $\zeta \leq \frac{2}{3}(\zeta + \frac{l_k}{4})$. $\text{Var}\left[Z_i | E_{k,p,q}\right] = \text{Var}\left[A_i^{(u)} \mid E_{k,p,q}\right] + \text{Var}\left[A_i^{(m)} \mid E_{k,p,q}\right] \overset{(a)}{\leq} \mathbb{E}\left[A_i^{(u)} \mid E_{k,p,q}\right] +$

$\mathbb{E}\left[A_i^{(m)} \mid E_{k,p,q}\right] = f(\zeta+\frac{l_k}{4}) + f(\zeta) \overset{(b)}{\leq} 2f(\zeta+\frac{l_k}{4})$ where (a) follows by $A_i \in \{0,1\}$ and (b) follows by the monotonicity of f . Thus, on event $E_{k,p,q}$, by Lemma 5.11, if we set η sufficiently large (independent of l_k, ε, δ), then with probability at least $1-\frac{\delta}{4\log\frac{1}{2\varepsilon}}$ CheckSignificant-Var$\left(\{Z_i\}_{i=1}^{N_k}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right)$ in Procedure 8 returns true.

Similarly, we can show that on event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$, by Lemma 5.11, with probability at least $1-\frac{\delta}{4\log\frac{1}{2\varepsilon}}$, CheckSignificant-Var$\left(\left\{A_i^{(v)} - A_i^{(m)}\right\}_{i=1}^{N_k}, \frac{\delta}{4\log\frac{1}{2\varepsilon}}\right)$ returns true.

Therefore, the loop at Line 6 will terminate after $n = N_k$ with probability at least $1-\frac{\delta}{4\log\frac{1}{2\varepsilon}}$ on event $E_{k,p,q}$. Therefore, with probability at least $1-\delta$, the number of samples queried is at most $\sum_{k=0}^{\log\frac{1}{2\varepsilon}-1} \frac{1}{f((\frac{3}{4})^k/4)} \ln\frac{\ln 1/\varepsilon}{\delta} = O\left(\frac{1}{f(\varepsilon/2)}\ln\frac{1}{\varepsilon}\left(\ln\frac{1}{\delta} + \ln\ln\frac{1}{\varepsilon}\right)\right)$. □

The query complexity given by Theorem 5.14 is independent of β that decides the flipping rate, and consequently smaller than the bound in Theorem 5.13. This improvement is due to the use of abstention responses, which become much more informative under Condition 3.

5.4.3 Lower Bounds

In this subsection, we give lower bounds of query complexity in the one-dimensional case and establish near optimality of Algorithm 7. We will give corresponding lower bounds for the high-dimensional case in the next section.

First, we introduce some notations for this section. Given a labeler L and an active learning algorithm $\mathcal{A}$, denote by $P_{L,\mathcal{A}}^n$ the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_L(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$. We will drop the subscripts from $P_{L,\mathcal{A}}^n$ and $P_L(Y|X)$ when it is clear from the context.

We will use Fano's method shown as below to prove the lower bounds.

Lemma 5.15. *Let Θ be a class of parameters, and $\{P_\theta : \theta \in \Theta\}$ be a class of probability distributions indexed by Θ over some sample space $\mathcal{X}$. Let $d : \Theta \times \Theta \to \mathbb{R}$ be a semi-metric. Let $\mathcal{V} = \{\theta_1, \dots, \theta_M\} \subseteq \Theta$ such that $\forall i \neq j$, $d(\theta_i, \theta_j) \geq 2s > 0$. Let $\bar{P} = \frac{1}{M}\sum_{\theta \in \mathcal{V}} P_\theta$. If $d_{KL}\left(P_\theta \parallel \bar{P}\right) \leq \delta$ for any $\theta \in \mathcal{V}$, then for any algorithm $\hat{\theta}$ that given a sample X drawn from P_θ outputs $\hat{\theta}(X) \in \Theta$, the following inequality holds:*

$$\sup_{\theta \in \Theta} P_\theta\left(d(\theta, \hat{\theta}(X)) \geq s\right) \geq 1 - \frac{\delta + \ln 2}{\ln M}$$

Proof. For any algorithm $\hat{\theta}$, define a test function $\hat{\Psi} : \mathcal{X} \to \{1, \dots, M\}$ such that $\hat{\Psi}(X) = \arg\min_{i \in \{1,\dots,M\}} d(\hat{\theta}(X), \theta_i)$. We have

$$\sup_{\theta \in \Theta} P_\theta\left(d(\theta, \hat{\theta}(X)) \geq s\right) \geq \max_{\theta \in \mathcal{V}} P_\theta\left(d(\theta, \hat{\theta}(X)) \geq s\right) \geq \max_{i \in \{1,\dots,M\}} P_{\theta_i}\left(\hat{\Psi}(X) \neq i\right)$$

Let V be a random variable uniformly taking values from $\mathcal{V}$, and X be drawn from P_V. By Fano's Inequality, for any test function $\Psi : \mathcal{X} \to \{1, \dots, M\}$

$$\max_{i \in \{1,\dots,M\}} P_{\theta_i}\left(\Psi(X) \neq i\right) \geq 1 - \frac{I(V;X) + \ln 2}{\ln M}$$

The desired result follows by the fact that $I(V;X) = \frac{1}{M}\sum_{\theta \in \mathcal{V}} d_{\mathrm{KL}}\left(P_\theta \parallel \bar{P}\right)$. □

Our query complexity (Theorem 5.14) for the algorithm is also almost tight under Conditions 1 and 3 with a polynomial abstention rate.

Theorem 5.16. *There is a universal constant $\delta_0 \in (0,1)$ and a labeler L satisfying Conditions 1, 2, and 3 with $f(x) = C'x^{\alpha}$ ($C' > 0$ and $0 < \alpha \leq 2$ are constants), such that for any active learning algorithm $\mathcal{A}$, there is a $\theta^* \in [0,1]$, such that for small enough ε, $\Lambda(\varepsilon, \delta_0, \mathcal{A}, L, \theta^*) \geq \Omega\left(\varepsilon^{-\alpha}\right)$.*

Proof of Theorem 5.16. [2] Without lose of generality, let $C = C' = 1$ (C is defined in Condition 2). Let $\varepsilon \leq \frac{1}{4}\min\left\{\left(\frac{1}{2}\right)^{1/\beta}, \left(\frac{4}{5}\right)^{1/\alpha}, \frac{1}{4}\right\}$. We will prove the desired result using Lemma 5.15.

First, we construct $\mathcal{V}$ and P_{θ}. For any $k \in \{0,1,2,3\}$, let $P_{L_k}(Y \mid X)$ be the distribution of the labeler L_k's response with the ground truth $\theta_k = k\varepsilon$:

$$
\begin{aligned}
P_{L_k}\left(Y = \perp | x\right) &= 1 - \left| x - \frac{1}{2} - k\varepsilon \right|^{\alpha} \\
P_{L_k}\left(Y = 0 | x\right) &= \begin{cases} \left(x - \frac{1}{2} - k\varepsilon\right)^{\alpha}\left(1 - \left(x - \frac{1}{2} - k\varepsilon\right)^{\beta}\right)/2 & x > \frac{1}{2} + k\varepsilon \\ \left(\frac{1}{2} + k\varepsilon - x\right)^{\alpha}\left(1 + \left(\frac{1}{2} + k\varepsilon - x\right)^{\beta}\right)/2 & x \leq \frac{1}{2} + k\varepsilon \end{cases} \\
P_{L_k}\left(Y = 1 | x\right) &= \begin{cases} \left(x - \frac{1}{2} - k\varepsilon\right)^{\alpha}\left(1 + \left(x - \frac{1}{2} - k\varepsilon\right)^{\beta}\right)/2 & x > \frac{1}{2} + k\varepsilon \\ \left(\frac{1}{2} + k\varepsilon - x\right)^{\alpha}\left(1 - \left(\frac{1}{2} + k\varepsilon - x\right)^{\beta}\right)/2 & x \leq \frac{1}{2} + k\varepsilon \end{cases}
\end{aligned}
$$

Clearly, P_{L_k} complies with Conditions 1, 2 and 3.

Define P_k^n to be the distribution of n samples $\left\{(X_i, Y_i)\right\}_{i=1}^{n}$ where Y_i is drawn from distribution $P_{L_k}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\left\{(X_j, Y_j)\right\}_{j=1}^{i-1}$.

Define $\bar{P}_L = \frac{1}{4}\sum_j P_{L_j}$ and $\bar{P}^n = \frac{1}{4}\sum_j P_k^n$. We take Θ to be $[0,1]$, and $d(\theta_1, \theta_2) = |\theta_1 - \theta_2|$

[2] Actually we can use Le Cam's method to prove this one-dimensional case (which only needs to construct 2 distributions instead of 4 here), but this proof can be generalized to the multidimensional case more easily.

in Lemma 5.15. To use Lemma 5.15, we need to bound $d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}^n\right)$ for $k \in \{0,1,2,3\}$.

For any $k \in \{0,1,2,3\}$,

$$\begin{aligned}
& d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}_0^n\right) \\
= & \mathbb{E}_{P_k^n}\left(\ln \frac{P_k^n\left(\left\{(X_i,Y_i)\right\}_{i=1}^n\right)}{\bar{P}^n\left(\left\{(X_i,Y_i)\right\}_{i=1}^n\right)}\right) \\
= & \mathbb{E}_{P_k^n}\left(\ln \frac{P_k^n\left(X_1\right) P_k^n\left(Y_1 \mid X_1\right) P_k^n\left(X_2 \mid X_1,Y_1\right) \cdots P_k^n\left(Y_n \mid X_1,Y_1,\ldots,X_n\right)}{\bar{P}^n\left(X_1\right) \bar{P}^n\left(Y_1 \mid X_1\right) \bar{P}^n\left(X_2 \mid X_1,Y_1\right) \cdots \bar{P}^n\left(Y_n \mid X_1,Y_1,\ldots,X_n\right)}\right) \\
\overset{(a)}{=} & \mathbb{E}_{P_k^n}\left(\ln \frac{\Pi_{i=1}^n P_{L_k}\left(Y_i|X_i\right)}{\Pi_{i=1}^n \bar{P}_L\left(Y_i|X_i\right)}\right) \\
= & \sum_{i=1}^n \mathbb{E}_{P_k^n}\left(\mathbb{E}_{P_k^n}\left(\ln \frac{P_{L_k}\left(Y_i|X_i\right)}{\bar{P}_L\left(Y_i|X_i\right)} \mid X^n\right)\right) \\
\leq & n \max_{x\in[0,1]} d_{\mathrm{KL}}\left(P_{L_k}(Y \mid x) \parallel \bar{P}_L(Y \mid x)\right)
\end{aligned} \tag{5.2}$$

(a) follows by the fact that $P_k^n\left(X_{i+1} \mid X_1,Y_1,\ldots X_i,Y_i\right) = \bar{P}^n\left(X_{i+1} \mid X_1,Y_1,\ldots,X_i,Y_i\right)$ since X_{i+1} is drawn by the same algorithm based solely on the knowledge of $\left\{(X_j,Y_j)\right\}_{j=1}^i$ regardless of the labeler's response distribution, and that $P_k^n\left(Y_i \mid X_1,Y_1,\ldots,X_i\right) = P_{\mathcal{L}_k}\left(Y_i|X_i\right)$ and $\bar{P}^n\left(Y_i \mid X_1,Y_1,\ldots,X_i\right) = \bar{P}_L\left(Y_i|X_i\right)$ by definition.

For any $k \in \{1,2,3\}, x \in [0,1]$,

$$\bar{P}_L(\cdot \mid x) \geq \frac{P_{L_0}(\cdot \mid x) + P_{L_k}(\cdot \mid x)}{4} \tag{5.3}$$

For any $k \in \{0,1,2,3\}, x \in [0,1], y \in \{1,-1,\perp\}$

$$
\begin{aligned}
&\left(\bar{P}_L(Y=y\mid x)-P_{L_k}(Y=y\mid x)\right)^2\\
=&\left(\sum_j \frac{1}{4}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)+\left(P_{L_0}(Y=y\mid x)-P_{L_k}(Y=y\mid x)\right)\right)^2\\
\leq&\left(\frac{5}{16}\sum_{j>0}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2+5\left(P_{L_0}(Y=y\mid x)-P_{L_k}(Y=y\mid x)\right)^2\right)\\
\leq&6\sum_{j>0}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2 \qquad (5.4)
\end{aligned}
$$

where the first inequality follows by $\left(\sum_{i=0}^4 a_i\right)^2 \leq 5\sum_{i=0}^4 a_i^2$ by letting $a_j=\frac{1}{4}(P_{L_j}(Y= y\mid x)-P_{L_0}(Y=y\mid x))$ for $j=0,\ldots,3$ and $a_4=P_{L_0}(Y=y\mid x)-P_{L_k}(Y=y\mid x)$, and noting that $a_0=0$ under this setting.

Thus,

$$
\begin{aligned}
&d_{\mathrm{KL}}\left(P_{L_k}(Y\mid x)\parallel \bar{P}_L(Y\mid x)\right)\\
\leq&\sum_y \frac{1}{\bar{P}_L(Y=y\mid x)}\left(P_{L_k}(Y=y\mid x)-\bar{P}_L(Y=y\mid x)\right)^2\\
\leq&24\sum_{j>0}\sum_y \frac{1}{P_{L_j}(y\mid x)+P_{L_0}(y\mid x)}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2\\
\leq&O(\varepsilon^{\alpha})
\end{aligned}
$$

The first inequality follows from Lemma B.5. The second inequality follows by (5.3) and (5.4). The last inequality follows by applying Lemma B.6 to $P_{L_0}(\cdot\mid x)$ and $P_{L_j}(\cdot\mid x)$ and the assumption $\alpha\leq 2$.

Therefore, we have $d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}_0^n\right) = nO(\varepsilon^{\alpha})$. By setting $n = \varepsilon^{-\alpha}$, we get $d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}_0^n\right) \leq O(1)$, and thus by Lemma 5.15,

$$\sup_{\theta} P_{\theta}\left(d(\theta, \hat{\theta}(X)) \geq \Omega(\varepsilon)\right) \geq 1 - \frac{O(1) + \ln 2}{\ln 4} = O(1)$$

□

5.4.4 Remarks

Our results confirm the intuition that learning with abstention is easier than learning with noisy labels. This is true because a noisy label might mislead the learning algorithm, but an abstention response never does. Our analysis shows, in particular, that if the labeler never abstains, and outputs completely noisy labels with probability bounded by $1 - |x - \theta^*|^{\gamma}$ (i.e., $P(Y \neq \mathbb{1}[x > \theta^*] \mid x) \leq \frac{1}{2}\left(1 - |x - \theta^*|^{\gamma}\right)$), then the near optimal query complexity of $\tilde{O}\left(\varepsilon^{-2\gamma}\right)$ is significantly larger than the near optimal $\tilde{O}\left(\varepsilon^{-\gamma}\right)$ query complexity associated with a labeler who only abstains with probability $P(Y = \perp \mid x) \leq 1 - |x - \theta^*|^{\gamma}$ and never flips a label. More precisely, while in both cases the labeler outputs the same amount of corrupted labels, the query complexity of the abstention-only case is significantly smaller than the noise-only case.

Note that the query complexity of Algorithm 7 consists of two kinds of queries: queries which return 0/1 labels and are used by function CheckSignificant, and queries which return abstention and are used by function CheckSignificant-Var. Algorithm 7 will stop querying when the responses of one of the two kinds of queries are statistically significant. Under Condition 2, our proof actually shows that the optimal number of queries is dominated by the number of queries used by CheckSignificant function. In other words, a simplified variant of Algorithm 7 which excludes use of abstention feedback is near optimal. Similarly, under Condition 3, the optimal query complexity is dominated by the number of queries used by CheckSignificant-Var function. Hence the variant of Algorithm 7 which disregards 0/1 labels would be near optimal.

5.5 The Multidimensional Case

We follow [CN08] to generalize the results from one-dimensional thresholds to the d-dimensional ($d > 1$) smooth boundary fragment class $\Sigma(K,\gamma)$.

5.5.1 Lower bounds

Theorem 5.17. *There are universal constants $\delta_0 \in (0,1)$, $c_0 > 0$, and a labeler L satisfying Conditions 1 and 2, such that for any active learning algorithm $\mathcal{A}$, there is a $g^* \in \Sigma(K,\gamma)$, such that for small enough ε, $\Lambda(\varepsilon,\delta_0,\mathcal{A},L,g^*) \geq \Omega\left(\frac{1}{f(c_0\varepsilon)}\varepsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$.*

Again, we will use Lemma 5.15 to prove the lower bounds for d-dimensional cases. We first construct $\{P_\theta : \theta \in \Theta\}$ using a similar idea with [CN08], and then use Lemma B.7 to select a subset $\tilde{\Theta} \subset \Theta$ to apply Lemma 5.15.

Proof of Theorem 5.17. Again, without lose of generality, let $C = 1$. Recall that we have defined $\tilde{x}$ to be $(x_1,\ldots,x_{d-1})$ for $x = (x_1,\ldots,x_d) \in \mathbb{R}^d$. Define $m = \left(\frac{1}{\varepsilon}\right)^{1/\gamma}$. $\mathcal{L} = \left\{0,\frac{1}{m},\ldots,\frac{m-1}{m}\right\}^{d-1}$, $h(\tilde{x}) = \Pi_{i=1}^{d-1}\exp\left(-\frac{1}{1-4x_i^2}\right)\mathbb{1}\left\{|x_i| < \frac{1}{2}\right\}$, $\phi_l(\tilde{x}) = Km^{-\gamma}h(m(\tilde{x}-l)-\frac{1}{2})$ where $l \in \mathcal{L}$. It is easy to check $\phi_l(\tilde{x})$ is (K,γ)-Hölder smooth and has bounded support $[l_1,l_1+\frac{1}{m}]\times\cdots\times[l_{d-1},l_{d-1}+\frac{1}{m}]$, which implies that for different $l_1,l_2 \in \mathcal{L}$, the support of ϕ_{l_1} and ϕ_{l_2} do not intersect.

Let $\Omega = \{0,1\}^{m^{d-1}}$. For any $\omega \in \Omega$, define $g_\omega(\tilde{x}) = \sum_{l\in\mathcal{L}}\omega_l\phi_l(\tilde{x})$. For each $\omega \in \Omega$, define the conditional distribution of labeler L_ω's response as follows:

For $x_d \leq A$, $P_{L_\omega}(y = \perp | x) = 1 - f(A)$, and $P_{L_\omega}(y \neq \mathbb{1}(x_d > g_\omega(\tilde{x}))|x, y \neq \perp) = \frac{1}{2}(1 - \left|x_d - g_\omega(\tilde{x})\right|^\beta)$;

For $x_d \geq A$, $P_{L_\omega}(y = \perp | x) = 1 - f(x_d)$, and $P_{L_\omega}(y \neq \mathbb{1}(x_d > g_\omega(\tilde{x}))|x, y \neq \perp) = \frac{1}{2}(1 - x_d^\beta)$.

Here, $A = c \max \phi(\tilde{x}) = c'\varepsilon$ for some constants c, c'.

It can be easily verified that P_{L_ω} satisfies Conditions 1 and 2. Note that $g_\omega(\tilde{x})$ can be seen as the underlying decision boundary for labeler P_{L_ω}.

Define P_ω^n to be the distribution of n samples $\left\{(X_i, Y_i)\right\}_{i=1}^n$ where Y_i is drawn from distribution $P_{L_\omega}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\left\{(X_j, Y_j)\right\}_{j=1}^{i-1}$.

By Lemma B.7, when ε is small enough so that m^{d-1} is large enough, there is a subset $\left\{\omega^{(1)}, \ldots, \omega^{(M)}\right\} \subset \Omega$ such that $\left\|\omega^{(i)} - \omega^{(j)}\right\|_0 \geq m^{d-1}/12$ for any $0 \leq i < j \leq M$ and $M \geq 2^{m^{d-1}/48}$. Define $P_i^n = P_{\omega^{(i)}}^n, \bar{P}^n = \frac{1}{M}\sum_{i=1}^M P_i^n$.

Next, we will apply Lemma 5.15 to $\left\{\omega^{(1)}, \ldots, \omega^{(M)}\right\}$ with $d(\omega^{(i)}, \omega^{(j)}) = \left\|g_{\omega^{(i)}} - g_{\omega^{(j)}}\right\|$. We will lower-bound $d(\omega^{(i)}, \omega^{(j)})$ and upper-bound $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}^n\right)$.

For any $1 \leq i < j \leq M$,

$$\begin{aligned}
&\left\|g_{\omega^{(i)}} - g_{\omega^{(j)}}\right\| \\
=&\sum_{l \in \{1,\ldots,m\}^{d-1}} \left|\omega_l^{(i)} - \omega_l^{(j)}\right| K m^{-\gamma-(d-1)} \|h\| \\
\geq& m^{d-1}/12 * K m^{-\gamma-(d-1)} \|h\| \\
=& K m^{-\gamma} \|h\| / 12 \\
=& \Theta(\varepsilon)
\end{aligned}$$

By the convexity of KL-divergence, $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}^n\right) \leq \frac{1}{M}\sum_{j=1}^{M} d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right)$, so it suffices to upper-bound $d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right)$ for any i, j.

For any $1 < i, j \leq M$,

$$\begin{aligned}
&d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right)\\
\leq& n \max_{x\in[0,1]^d} d_{\mathrm{KL}}\left(P^n_{L_{\omega^{(i)}}}(Y \mid x) \parallel P^n_{L_{\omega^{(j)}}}(Y \mid x)\right)\\
=& n \max_{x\in[0,1]^d} P^n_{L_{\omega^{(i)}}}(Y \neq \perp \mid x) d_{\mathrm{KL}}\left(P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp) \parallel P^n_{L_{\omega^{(j)}}}(Y \mid x, Y \neq \perp)\right)
\end{aligned}$$

The inequality follows as (5.2) in the proof of Theorem 5.16. The equality follows since $P_\omega(y = \perp |x)$ is the same for all $\omega \in \Omega$.

If $x_d \geq A$, then $P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp) = P^n_{L_{\omega^{(j)}}}(Y \mid x, Y \neq \perp)$, so

$$d_{\mathrm{KL}}\left(P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp) \parallel P^n_{L_{\omega^{(j)}}}(Y \mid x, Y \neq \perp)\right) = 0.$$

If $x_d < A$, then $P^n_{L_{\omega^{(i)}}}(Y \neq \perp \mid x) = f(A)$. Therefore,

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq nf(A) \max_{x\in[0,1]^d} d_{\mathrm{KL}}\left(P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp) \parallel P^n_{L_{\omega^{(j)}}}(Y \mid x, Y \neq \perp)\right)$$

.

Apply Lemma B.5 to $P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp)$ and $P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp)$, and noting they are bounded above by a constant, we have $\max_{x\in[0,1]^d} d_{\mathrm{KL}}(P^n_{L_{\omega^{(i)}}}(Y \mid x, Y \neq \perp) \parallel P^n_{L_{\omega^{(j)}}}(Y \mid x, Y \neq \perp)) = O\left(A^{2\beta}\right)$. Thus,

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq nf(A)O\left(A^{2\beta}\right) = nf(c'\varepsilon)O(\varepsilon^{2\beta})$$

By setting $n = \frac{1}{f(c'\varepsilon)}\varepsilon^{-2\beta-\frac{d-1}{\gamma}}$, we get $d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq O\left(\varepsilon^{-\frac{d-1}{\gamma}}\right)$. The desired results follows by Lemma 5.15. □

Theorem 5.18. *There is a universal constant $\delta_0 \in (0,1)$ and a labeler L satisfying Conditions 1, 2, and Condition 3 with $f(x) = C'x^{\alpha}$ ($C' > 0$ and $0 < \alpha \leq 2$ are constants), such that for any active learning algorithm $\mathcal{A}$, there is a $g^* \in \Sigma(K,\gamma)$, such that for small enough ε, $\Lambda(\varepsilon, \delta_0, \mathcal{A}, L, g^*) \geq \Omega\left(\varepsilon^{-\alpha-\frac{d-1}{\gamma}}\right)$.*

The proof of Theorem 5.18 follows the same structure.

Proof of Theorem 5.18. As in the proof of Theorem 5.17, let $C = C' = 1$, and define $m = \left(\frac{1}{\varepsilon}\right)^{1/\gamma}$. $\mathcal{L} = \left\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\right\}^{d-1}$, $h(\tilde{x}) = \Pi_{i=1}^{d-1}\exp\left(-\frac{1}{1-4x_i^2}\right)\mathbb{1}\left\{|x_i| < \frac{1}{2}\right\}$, $\phi_l(\tilde{x}) = Km^{-\gamma}h(m(\tilde{x}-l)-\frac{1}{2})$ where $l \in \mathcal{L}$. Let $\Omega = \{0,1\}^{m^{d-1}}$. For any $\omega \in \Omega$, define $g_\omega(\tilde{x}) = \frac{1}{2} + \sum_{l\in\mathcal{L}}\omega_l\phi_l(\tilde{x})$, which can be seen as a decision boundary. $A = \max\phi(\tilde{x}) = c'\varepsilon$ for some constants c'.

Let $g_+(\tilde{x}) = g_{(1,1,\ldots,1)}(\tilde{x}) = \sum_{l\in\mathcal{L}}\phi_l(\tilde{x})$, $g_-(\tilde{x}) = g_{(0,0,\ldots,0)}(\tilde{x}) = 0$. In other words, g_+ is the "highest" boundary, and g_- is the "lowest" boundary.

For each $\omega \in \Omega$, define the conditional distribution of labeler L_ω's response as follows:

$$P_{L_\omega}(y = \perp | x) = 1 - \left|x_d - g_\omega(\tilde{x})\right|^{\alpha}$$

$$P_{L_\omega}(y \neq \mathbb{1}(x_d > g_\omega(\tilde{x}))|x, y \neq \perp) = \frac{1}{2}\left(1 - \left|x_d - g_\omega(\tilde{x})\right|^{\beta}\right)$$

It can be easily verified that P_{L_ω} satisfies Conditions 1, 2, and 3.

Let $P_+(\cdot \mid x) = P_{L_{(1,1,\ldots,1)}}(\cdot \mid x)$, $P_-(\cdot \mid x) = P_{L_{(0,0,\ldots,0)}}(\cdot \mid x)$. By the construction of g, for any

$x \in [0,1]^d$, any $\omega \in \Omega$, $P_{L_\omega}(\cdot \mid x)$ equals either $P_+(\cdot \mid x)$ or $P_-(\cdot \mid x)$.

Define P_ω^n to be the distribution of n samples $\left\{(X_i, Y_i)\right\}_{i=1}^n$ where Y_i is drawn from distribution $P_{L_\omega}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\left\{(X_j, Y_j)\right\}_{j=1}^{i-1}$.

By Lemma B.7, when ε is small enough so that m^{d-1} is large enough,, there is a subset $\Omega' = \left\{\omega^{(1)}, \ldots, \omega^{(M)}\right\} \subset \Omega$ such that (i) (well-separated) $\left\|\omega^{(i)} - \omega^{(j)}\right\|_0 \geq m^{d-1}/12$ for any $0 \leq i < j \leq M$, $M \geq 2^{m^{d-1}/48}$; and (ii) (well-balanced) for any $j = 1, \ldots, m^{d-1}$, $\frac{1}{24} \leq \frac{1}{M}\sum_{i=1}^M \omega_j^{(i)} \leq \frac{3}{24}$.

Define $P_i^n = P_{\omega^{(i)}}^n, \bar{P}^n = \frac{1}{M}\sum_{i=1}^M P_i^n$. Define $P_{L_i} = P_{L_{\omega^{(i)}}}$, $\bar{P}_L = \frac{1}{M}\sum_{i=1}^M P_{L_i}$. By the well-balanced property, for any $x \in [0,1]^d$, $\bar{P}_L(\cdot \mid x)$ is between $\frac{1}{24}P_+(\cdot \mid x) + \frac{23}{24}P_-(\cdot \mid x)$ and $\frac{3}{24}P_+(\cdot \mid x) + \frac{21}{24}P_-(\cdot \mid x)$. Therefore

$$\bar{P}_L(\cdot \mid x) \geq \frac{1}{24}\left(P_+(\cdot \mid x) + P_-(\cdot \mid x)\right) \tag{5.5}$$

Moreover, since $P_{L_i}(\cdot \mid x)$ can only take $P_+(\cdot \mid x)$ or $P_-(\cdot \mid x)$ for any x,

$$\left|P_{L_i}(\cdot \mid x) - \bar{P}_L(\cdot \mid x)\right| \leq \left|P_+(\cdot \mid x) - P_-(\cdot \mid x)\right| \tag{5.6}$$

Next, we will apply Lemma 5.15 to $\left\{\omega^{(1)}, \ldots, \omega^{(M)}\right\}$ with $d(\omega^{(i)}, \omega^{(j)}) = \left\|g_{\omega^{(i)}} - g_{\omega^{(j)}}\right\|$. We already know from the proof of Theorem 5.17 $\left\|g_{\omega^{(i)}} - g_{\omega^{(j)}}\right\| = \Omega(\varepsilon)$.

For any $0 < i \leq M$, $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}_0^n\right) \leq n \max_{x \in [0,1]^d} d_{\mathrm{KL}}\left(P_{L_i}(Y \mid x) \parallel \bar{P}_L(Y \mid x)\right)$. For any

$x \in [0,1]^d$,

$$
\begin{aligned}
& d_{\mathrm{KL}}\left(P_{L_i}(Y \mid x) \parallel \bar{P}_L(Y \mid x)\right) \\
\leq & \sum_y \frac{1}{\bar{P}_L(Y = y \mid x)} \left(P_{L_i}(Y = y \mid x) - \bar{P}_L(Y = y \mid x)\right)^2 \\
\leq & \sum_y \frac{24}{P_+(y \mid x) + P_-(y \mid x)} \left(P_+(Y = y \mid x) - P_-(Y = y \mid x)\right)^2 \\
\leq & O(A^\alpha)
\end{aligned}
$$

The first inequality follows from Lemma B.5. The second inequality follows by (5.5) and (5.6). The last inequality follows by applying Lemma B.6 to $P_+(\cdot \mid x)$ and $P_-(\cdot \mid x)$, setting the ε in Lemma B.6 to be $g_\omega(\tilde{x})$, and using $g_\omega(\tilde{x}) \leq A$ and the assumption $\alpha \leq 2$.

Therefore, we have

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_0^n\right) \leq nO\left(A^\alpha\right) = nO(\varepsilon^\alpha)$$

By setting $n = \varepsilon^{-\alpha - \frac{d-1}{\gamma}}$, we get $d_{\mathrm{KL}}\left(P_i^n \parallel P_0^n\right) \leq O\left(\varepsilon^{-\frac{d-1}{\gamma}}\right)$. Thus by Lemma 5.15,

$$\sup_\theta P_\theta\left(d(\theta, \hat{\theta}(X)) \geq \Omega(\varepsilon)\right) \geq 1 - \frac{O\left(\varepsilon^{-\frac{d-1}{\gamma}}\right) + \ln 2}{\varepsilon^{-\frac{d-1}{\gamma}}/48} = O(1)$$

, from which the desired result follows. □

Algorithm 9 The active learning algorithm for the smooth boundary fragment class

```
Input: $\delta$, $\varepsilon$, $\gamma$
$M \leftarrow \Theta\left(\varepsilon^{-1/\gamma}\right)$. $\mathcal{L} \leftarrow \left\{\frac{0}{M}, \frac{1}{M}, \dots, \frac{M-1}{M}\right\}^{d-1}$
For each $l \in \mathcal{L}$, apply Algorithm 7 with parameter ($\varepsilon$, $\delta/M^{d-1}$) to learn a threshold $g_l$ that approximates $g^*(l)$
Partition the instance space into cells $\left\{I_q\right\}$ indexed by $q \in \left\{0, 1, \dots, \frac{M}{\gamma} - 1\right\}^{d-1}$, where
    $I_q = \left[\frac{q_1\gamma}{M}, \frac{(q_1+1)\gamma}{M}\right] \times \cdots \times \left[\frac{q_{d-1}\gamma}{M}, \frac{(q_{d-1}+1)\gamma}{M}\right]$
For each cell $I_q$, perform a polynomial interpolation: $g_q(\tilde{x}) = \sum_{l \in I_q \cap \mathcal{L}} g_l Q_{q,l}(\tilde{x})$, where
    $Q_{q,l}(\tilde{x}) = \prod_{i=1}^{d-1} \prod_{j=0, j \neq Ml_i - \gamma q_i}^{\gamma} \frac{\tilde{x}_i - (\gamma q_i + j)/M}{l_i - (\gamma q_i + j)/M}$
Output: $g(\tilde{x}) = \sum_{q \in \left\{0,1,\dots,\frac{M}{\gamma}-1\right\}^{d-1}} g_q(\tilde{x}) \mathbb{1}\left[\tilde{x} \in q\right]$
```

5.5.2 Algorithm and Analysis

Recall the decision boundary of the smooth boundary fragment class can be seen as the epigraph of a smooth function $[0,1]^{d-1} \to [0,1]$. For $d > 1$, we can reduce the problem to the one-dimensional problem by discretizing the first $d-1$ dimensions of the instance space and then perform a polynomial interpolation. The algorithm is shown as Algorithm 9. For the sake of simplicity, we assume γ, M/γ in Algorithm 9 are integers.

We have similar consistency guarantee and upper bounds as in the one-dimensional case.

Theorem 5.19. *Let g^* be the ground truth. If the labeler L satisfies Condition 1 and Algorithm 9 stops to output g, then $\|g^* - g\| \leq \varepsilon$ with probability at least $1 - \frac{\delta}{2}$.*

Theorem 5.20. *Let g^* be the ground truth, and g be the output of Algorithm 9. Under Conditions 1 and 2, with probability at least $1-\delta$, Algorithm 9 makes at most $\tilde{O}\left(\frac{d}{f\left(\varepsilon/2\right)}\varepsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$ queries.*

Theorem 5.21. *Let g^* be the ground truth, and g be the output of Algorithm 9. Under Conditions 1* *and 3, with probability at least* $1-\delta$*, Algorithm 9 makes at most* $\tilde{O}\left(\frac{d}{f(\varepsilon/2)}\varepsilon^{-\frac{d-1}{\gamma}}\right)$ *queries.*

To prove the d-dimensional case, we only need to use a union bound to show that with high probability all calls of Algorithm 7 succeed, and consequently the output boundary g produced by polynomial interpolation is close to the true underlying boundary due to the smoothness assumption of g^*.

Proof of Theorem 5.19. For $q \in \left\{0,1,\ldots,\frac{M}{\gamma}-1\right\}^{d-1}$, define the "polynomial interpolation" version of g^* as

$$g_q^*(\tilde{x}) = \sum_{l \in I_q \cap \mathcal{L}} g^*(l) Q_{q,l}(\tilde{x})$$

Recall that we choose $M = O\left(\varepsilon^{-1/\gamma}\right)$.

By Theorem 5.12, each run of Algorithm 7 at the line 3 of Algorithm 9 will return a g_l such that $\left|g_l - g_q^*(l)\right| \leq \varepsilon$ with probability at least $1-\delta/2M^{d-1}$.

$$\begin{aligned}
&\left\|g-g^*\right\| \\
=& \sum_{q\in\{0,\ldots,M/\gamma-1\}^{d-1}} \left\|\left(g_q-g^*\right)\mathbb{1}\{\tilde{x}\in I_q\}\right\| \\
\leq& \sum_{q\in\{0,\ldots,M/\gamma-1\}^{d-1}} \left\|\left(g_q-g_q^*\right)\mathbb{1}\{\tilde{x}\in I_q\}\right\| + \left\|\left(g_q^*-g^*\right)\mathbb{1}\{\tilde{x}\in I_q\}\right\|
\end{aligned}$$

$$\begin{aligned}\left\|\left(g_q^* - g^*\right)\mathbb{1}\{\tilde{x}\in I_q\}\right\| &= \int_{I_q}\left|g_q^*(\tilde{x}) - g^*(\tilde{x})\right| d\tilde{x} \\ &= O\left(\int_{I_q} M^{-\gamma} d\tilde{x}\right) \\ &= O\left(M^{-\gamma-d+1}\right)\end{aligned}$$

The second equality follows from Lemma 3 of [CN08] that $\left|g_q(\tilde{x}) - g^*(\tilde{x})\right| = O\left(M^{-\gamma}\right)$ since g^* is γ-Hölder smooth.

$$\begin{aligned}&\left\|\left(g_q - g_q^*\right)\mathbb{1}\{\tilde{x}\in I_q\}\right\| \\ =&\sum_{l\in I_q\cap\mathcal{L}}\left|g_l - g_q^*(l)\right|\left\|Q_{q,l}\right\| \\ \leq&\sum_{l\in I_q\cap\mathcal{L}}\varepsilon\left\|Q_q\right\| \\ =&O(\varepsilon M^{-d+1})\end{aligned}$$

Therefore, overall we have $\|g - g^*\| \leq O\left(M^{-\gamma-d+1} + \varepsilon M^{-d+1}\right)\left(\frac{M}{\gamma}\right)^{d-1} = O(\varepsilon)$. □

Proof of Theorem 5.20. By Theorem 5.13, each run of Algorithm 7 at the line 3 of Algorithm 9 will make $\tilde{O}\left(\frac{d}{f(\varepsilon/2)}\varepsilon^{-2\beta}\right)$ queries with probability at least $1-\delta/M^{d-1}$, thus by a union bound, the total number of queries made is $\tilde{O}\left(\frac{d}{f(\varepsilon/2)}\varepsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$ with probability at least $1-\delta$. □

Proof of Theorem 5.21. The proof is similar to the previous proof. □

5.6 Acknowledgements

This chapter is based on the materials in Allerton Conference on Communication, Control and Computing 2015 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning from Noisy and Abstention Feedback”) [YCJ15] and Advances in Neural Information Processing Systems 2016 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning from Imperfect Labelers”) [YCJ16]. The dissertation author is the primary investigator and author of these materials.

Chapter 6

Active Learning with Logged Observational Data I: An Importance Sampling Solution

6.1 Introduction

In this chapter, we consider active learning with an auxiliary observational dataset. Counterfactual learning from observational data is an emerging problem that arises naturally in many applications. In this problem, the learner is given observational data – a set of examples selected according to some policy along with their labels – as well as access to the policy that selects the examples, and the goal is to construct a classifier with high performance on an entire population, not just the observational data distribution.

An example is predicting the efficacy of a treatment as a function of patient characteristics based on observed data. Doctors may assign the treatment to patients based on some predetermined rule; recording these patient outcomes produces a logged dataset where outcomes are observed conditioned on the doctors' assignment. A second example is recidivism prediction, where the goal is to predict whether a convict will re-offend. Judges use their own predefined policy to grant parole, and if parole is granted, then an outcome (reoffense or not) is observed. Thus the observed data records outcomes conditioned on the judges' parole policy, while the learner's goal is to learn a predictor over the entire population.

A major challenge in learning from logged data is that the logging policy may leave large areas of the data distribution under-explored. Consequently, empirical risk minimization (ERM) on the logged data leads to classifiers that may be highly suboptimal on the population. When the logging policy is known, a second option is to use a *weighted* ERM, that reweighs each observed labeled data point to ensure that it reflects the underlying population. However, this may lead to sample inefficiency if the logging policy does not adequately explore essential regions of the population. A final approach, typically used in clinical trials, is controlled random experimentation – essentially, ignore the logged data, and record outcomes for fresh examples drawn from the population. This approach is expensive due to the high cost of trials, and wasteful since it ignores the observed data.

Motivated by these challenges, we propose active learning to combine logged data with a small amount of strategically chosen labeled data that can be used to correct the bias in the logging policy. This solution has the potential to achieve the best of both worlds by limiting experimentation to achieve higher sample efficiency, and by making the most of the logged data. Specifically, we assume that in addition to the logged observational data, the learner has some additional unlabeled data that he can selectively ask an annotator to label. The learner's goal is to learn a highly accurate classifier over the entire population by using a combination of the logged data and with as few label queries to the annotator as possible.

How can we utilize logged data for better active learning? Prior work in this problem has looked at both probabilistic inference [SSS$^+$19, AZvdS19], and here we consider the standard classification setting. A naive approach is to use the logged data to come up with a *warm start* and then do standard active learning. In this work, we show that we can do even better. In addition to the warm start, we show how to use multiple importance sampling estimators to utilize the logged data more efficiently. Additionally, we introduce a novel sample selection bias correction technique that selectively avoids label queries for those examples that are highly represented in the logged data.

Combining these three approaches, we provide a new algorithm. We prove that our algorithm is statistically consistent, and has a lower label requirement than simple active learning that uses the logged data as a warm start. Finally, we evaluate our algorithm experimentally on various datasets and logging policies. Our experiments show that the performance of our method is either the best or close to the best for a variety of datasets and logging policies. This confirms that active learning to combine logged data with carefully chosen labeled data may indeed yield performance gains.

6.2 Setup

We are given a instance space $\mathcal{X}$, a label space $\mathcal{Y} = \{-1, +1\}$, and a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. Let D be an underlying data distribution over $\mathcal{X} \times \mathcal{Y}$. For simplicity, we assume $\mathcal{H}$ is a finite set, but our results can be generalized to VC-classes by standard arguments [VC71].

In the passive setting for learning with observational data, the learner has access to a logged observational dataset generated from the following process. First, m examples $\{(X_t, Y_t)\}_{t=1}^{m}$ are drawn i.i.d. from D. Then a logging policy $Q_0 : \mathcal{X} \to [0, 1]$ that describes the probability of observing the label is applied. In particular, for each example (X_t, Y_t) $(1 \leq t \leq m)$, an independent

Bernoulli random variable Z_t with expectation $Q_0(X_t)$ is drawn, and then the label Y_t is revealed to the learner if $Z_t = 1$[1]. We call $T_0 = \{(X_t, Y_t, Z_t)\}_{t=1}^{m}$ the logged dataset. We assume the learner knows the logging policy Q_0, and only observes instances $\{X_t\}_{t=1}^{m}$, indicators $\{Z_t\}_{t=1}^{m}$, and revealed labels $\{Y_t \mid Z_t = 1\}_{t=1}^{m}$.

In the active learning setting, in addition to the logged dataset, the learner has access to a stream of online data. In particular, there is a stream of additional n examples $\{(X_t, Y_t)\}_{t=m+1}^{m+n}$ drawn i.i.d. from distribution D. At time t ($m < t \leq m+n$), the learner applies a query policy to compute an indicator $Z_t \in \{0,1\}$, and then the label Y_t is revealed if $Z_t = 1$. The computation of Z_t may in general be randomized, and is based on the observed logged data T_0, previously observed instances $\{X_i\}_{i=m+1}^{t}$, decisions$\{Z_i\}_{i=m+1}^{t-1}$, and observed labels $\{Y_i \mid Z_i = 1\}_{i=m+1}^{t-1}$.

We focus on the active learning setting, and the goal of the learner is to learn a classifier $h \in \mathcal{H}$ from observed logged data and online data. Fixing D, Q_0, m, n, the performance is measured by: (1) the error rate $l(h) := \mathbb{P}_D(h(X) \neq Y)$ of the output classifier, and (2) the number of label queries on the online data. Note that the error rate is over the entire population D instead of conditioned on the logging policy, and that we assume the labels of the logged data T_0 come at no cost. In this work, we are interested in the situation where n, the size of the online stream, is smaller than m.

Notation Unless otherwise specified, all probabilities and expectations are over the draw of all random variables $\{(X_t, Y_t, Z_t)\}_{t=1}^{m+n}$. Define $q_0 = \inf_x Q_0(x)$. Define the optimal classifier $h^\star = \arg\min_{h \in \mathcal{H}} l(h)$, $\nu = l(h^\star)$. For any $r > 0, h \in \mathcal{H}$, define the $r-$ball around h as $B(h,r) = \left\{h' \in \mathcal{H} : \mathbb{P}(h(X) \neq h'(X)) \leq r\right\}$. For any $C \subseteq \mathcal{H}$, define the disagreement region $\text{DIS}(C) = \{x \in \mathcal{X} : \exists h_1 \neq h_2 \in C, h_1(X) \neq h_2(X)\}$.

[1] This generating process implies the standard unconfoundedness assumption in the counterfactual inference literature: $\mathbb{P}(Y_t, Z_t \mid X_t) = \mathbb{P}(Y_t \mid X_t)\mathbb{P}(Z_t \mid X_t)$. In other words, the label Y_t is conditionally independent with the action Z_t (indicating whether the label is observed) given the instance X_t.

6.3 Key Ideas

Our algorithm employs the disagreement-based active learning framework (Algorithm 1), but modifies the main DBAL algorithm in three key ways.

Key Idea 1: Warm-Start

Our algorithm applies a straightforward way of making use of the logged data T_0 inside the DBAL framework: to set the initial candidate set C_0 to be the set of classifiers that have a low empirical error on T_0.

Key Idea 2: Multiple Importance Sampling

Most learning algorithms, including DBAL, require estimating the error rate of a classifier. A good error estimator should be unbiased and of low variance. When instances are observed with different probabilities, a commonly used error estimator is the standard importance sampling estimator that reweighs each observed labeled example according to the inverse probability of observing it.

Consider a simplified setting where the logged dataset $T_0 = (X_i, Y_i, Z_i)_{i=1}^m$ and $\mathbb{P}(Z_i = 1 \mid X_i) = Q_0(X_i)$. On the online dataset $T_1 = (X_i, Y_i, Z_i)_{i=m+1}^{m+n}$, the algorithm uses a fixed query policy Q_1 to determine whether to query for labels, that is, $\mathbb{P}(Z_i = 1 \mid X_i) = Q_1(X_i)$ for $m < i \leq m+n$. Let $S = T_0 \cup T_1$.

In this setting, the standard importance sampling (IS) error estimator for a classifier h is:

$$l_{\text{IS}}(h,S) := \frac{1}{m+n}\sum_{i=1}^{m}\frac{\mathbb{1}\{h(X_i) \neq Y_i\}Z_i}{Q_0(X_i)} + \frac{1}{m+n}\sum_{i=m+1}^{m+n}\frac{\mathbb{1}\{h(X_i) \neq Y_i\}Z_i}{Q_1(X_i)}. \tag{6.1}$$

l_{IS} is unbiased, and its variance is proportional to $\sup_{i=0,1;x\in\mathcal{X}} \frac{1}{Q_i(x)}$. Although the learning algorithm can choose its query policy Q_1 to avoid $Q_1(X_i)$ to be too small for $i > m$, Q_0 is the logging policy that cannot be changed. When $Q_0(X_i)$ is small for some $i \leq m$, the estimator in (6.1) have a high variance such that it may be even better to just ignore the logged dataset T_0.

An alternative is the multiple importance sampling (MIS) estimator with balanced heuristic [VG95]:

$$l_{\mathrm{MIS}}(h,S) := \sum_{i=1}^{m+n} \frac{\mathbb{1}\{h(X_i) \neq Y_i\} Z_i}{mQ_0(X_i) + nQ_1(X_i)}. \tag{6.2}$$

It can be proved that $l_{\mathrm{MIS}}(h,S)$ is indeed an unbiased estimator for $l(h)$. Moreover, as proved in [OZ00, ABSJ17], (6.2) always has a lower variance than both (6.1) and the standard importance sampling estimator that ignores the logged data.

Thus, in our work, we use multiple importance sampling estimators instead of standard importance sampling estimators to which obtain a better performance guarantee.

We remark that the main purpose of using multiple importance sampling estimators here is to control the variance due to the predetermined logging policy. In the classical active learning setting without logged data, standard importance sampling can give satisfactory performance guarantees [BDL09, BHLZ10, HAH$^+$15].

Key Idea 3: A Sample Selection Bias Correction Query Strategy

The logging policy Q_0 introduces bias into the logged data: some examples may be underrepresented since Q_0 chooses to reveal their labels with lower probability. Our algorithm employs a sample selection bias correction query strategy to neutralize this effect. For any instance x in the online data, the algorithm would query for its label with a lower probability if $Q_0(x)$ is relatively large.

It is clear that a lower query probability leads to fewer label queries. Moreover, we claim that our sample selection bias correction strategy, though queries for less labels, does not deteriorate our theoretical guarantee on the error rate of the final output classifier. To see this, we note that we can establish a concentration bound for multiple importance sampling estimators that with probability at least $1-\delta$, for all $h \in \mathcal{H}$,

$$\begin{aligned} l(h) - l(h^\star) \leq & 2(l(h,S) - l(h^\star,S)) + \gamma_1 \sup_{x\in\mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{|\mathcal{H}|}{\delta}}{mQ_0(x) + nQ_1(x)} \\ & + \gamma_1 \sqrt{\sup_{x\in\mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{|\mathcal{H}|}{\delta}}{mQ_0(x) + nQ_1(x)} l(h^\star)} \end{aligned} \tag{6.3}$$

where m,n are sizes of logged data and online data respectively, Q_0 and Q_1 are query policy during the logging phase and the online phase respectively, and γ_1 is an absolute constant (see Corollary C.12 in Appendix for proof).

This concentration bound implies that for any $x \in \mathcal{X}$, if $Q_0(x)$ is large, we can set $Q_1(x)$ to be relatively small (as long as $mQ_0(x) + nQ_1(x) \geq \inf_{x'} mQ_0(x') + nQ_1(x')$) while achieving the same concentration bound. Consequently, the upper bound on the final error rate that we can establish from this concentration bound would not be impacted by the sample selection bias correction querying strategy.

One technical difficulty of applying both multiple importance sampling and the sample selection bias correction strategy to the DBAL framework is adaptivity. Applying both methods requires that the query policy and consequently the importance weights in the error estimator are updated with observed examples in each iteration. In this case, the summands of the error estimator are not independent, and the estimator becomes an adaptive multiple importance sampling estimator whose convergence property is still an open problem [CMMR12].

To circumvent this convergence issue and establish rigorous theoretical guarantees, in

Algorithm 10 Active learning with logged data

```
Input: confidence $\delta$, size of online data $n$, logging policy $Q_0$, logged data $T_0$.
$K \leftarrow \lceil \log n \rceil$.
$\tilde{S}_0 \leftarrow T_0^{(0)}$; $C_0 \leftarrow \mathcal{H}$; $D_0 \leftarrow \mathcal{X}$; $\xi_0 \leftarrow \inf_{x \in \mathcal{X}} Q_0(x)$.
for $k = 0, \dots, K-1$ do
    $\delta_k \leftarrow \frac{\delta}{(k+1)(k+2)}$; $\sigma(k,\delta) \leftarrow \frac{\log\frac{|\mathcal{H}|}{\delta}}{m_k\xi_k+n_k}$; $\Delta_k(h,h') \leftarrow \gamma_0\sigma(k,\frac{\delta_k}{2}) + \gamma_0\sqrt{\sigma(k,\frac{\delta_k}{2})\rho_{\tilde{S}_k}(h,h')}$.
    $\triangleright$ $\gamma_0$ is an absolute constant defined in Lemma C.13.
    $\hat{h}_k \leftarrow \arg\min_{h \in C_k} l(h, \tilde{S}_k)$.
    Define the candidate set
        $$C_{k+1} \leftarrow \{h \in C_k \mid l(h,\tilde{S}_k) \le l(\hat{h}_k,\tilde{S}_k) + \Delta_k(h,\hat{h}_k)\}$$
    and its disagreement region $D_{k+1} \leftarrow \mathrm{DIS}(C_{k+1})$.
    Define $\xi_{k+1} \leftarrow \inf_{x \in D_{k+1}} Q_0(x)$, and $Q_{k+1}(x) \leftarrow \mathbb{1}\{Q_0(x) \le \xi_{k+1} + 1/\alpha\}$.
    Draw $n_{k+1}$ samples $\{(X_t,Y_t)\}_{t=m+n_1\cdots+n_k+1}^{m+n_1+\cdots+n_{k+1}}$, and present $\{X_t\}_{t=m+n_1+\cdots+n_k+1}^{m+n_1+\cdots+n_{k+1}}$ to the algorithm.
    for $t = m+n_1+\cdots+n_k+1$ to $m+n_1+\cdots+n_{k+1}$ do
        $Z_t \leftarrow Q_{k+1}(X_t)$.
        if $Z_t = 1$ then
            If $X_t \in D_{k+1}$, query for label: $\tilde{Y}_t \leftarrow Y_t$; otherwise infer $\tilde{Y}_t \leftarrow \hat{h}_k(X_t)$.
        end if
    end for
    $\tilde{T}_{k+1} \leftarrow \{X_t,\tilde{Y}_t,Z_t\}_{t=m+n_1+\cdots+n_k+1}^{m+n_1+\cdots+n_{k+1}}$.
    $\tilde{S}_{k+1} \leftarrow T_0^{(k+1)} \cup \tilde{T}_{k+1}$.
end for
Output $\hat{h} = \arg\min_{h \in C_K} l(h,\tilde{S}_K)$.
```

each iteration, we compute the error estimator from a fresh sample set. In particular, we partition the logged data and the online data stream into disjoint subsets, and we use one logged subset and one online subset for each iteration.

6.4 Algorithm

The Algorithm is shown as Algorithm 10. Algorithm 10 runs in K iterations where $K = \lceil \log n \rceil$ (recall n is the size of the online data stream). For simplicity, we assume $n = 2^K - 1$.

As noted in the previous subsection, we require the algorithm to use a disjoint sample set for each iteration. Thus, we partition the data as follows. The online data stream is partitioned into K parts $T_1, \cdots, T_K$ of sizes $n_1 = 2^0, \cdots, n_K = 2^{K-1}$. We define $n_0 = 0$ for completeness. The logged data T_0 is partitioned into $K+1$ parts $T_0^{(0)}, \cdots, T_0^{(K)}$ of sizes $m_0 = m/3, m_1 = \alpha n_1, m_2 = \alpha n_2, \cdots, m_K = \alpha n_K$ (where $\alpha = 2m/3n$ and we assume $\alpha \geq 1$ is an integer for simplicity. m_0 can take other values as long as it is a constant factor of m). The algorithm uses $T_0^{(0)}$ to construct an initial candidate set, and uses $S_k := T_0^{(k)} \cup T_k$ in iteration k.

Algorithm 10 uses the disagreement-based active learning framework. At iteration k ($k = 0, \cdots, K-1$), it first constructs a candidate set C_{k+1} which is the set of classifiers whose training error (using the multiple importance sampling estimator) on $T_0^{(k)} \cup \tilde{T}_k$ is small, and its disagreement region D_{k+1}. At the end of the k-th iteration, it receives the $(k+1)$-th part of the online data stream $\{X_i\}_{i=m+n_1\cdots+n_k+1}^{m+n_1\cdots+n_{k+1}}$ from which it can query for labels. It only queries for labels inside the disagreement region D_{k+1}. For any example X outside the disagreement region, Algorithm 10 infers its label $\tilde{Y} = \hat{h}_k(X)$. Throughout this chapter, we denote by T_k, S_k the set of examples with original labels, and by $\tilde{T}_k$, $\tilde{S}_k$ the set of examples with inferred labels. The algorithm only observes $\tilde{T}_k$ and $\tilde{S}_k$.

Algorithm 10 uses aforementioned sample selection bias correction query strategy, which leads to fewer label queries than the standard disagreement-based algorithms. To simplify our analysis, we round the query probability $Q_k(x)$ to be 0 or 1.

6.5 Analysis

In this section, we establish theoretical guarantees for the proposed algorithm. All proofs are deferred to Appendix.

6.5.1 Consistency

We first introduce some additional quantities.

Define $h^\star := \min_{h \in \mathcal{H}} l(h)$ to be the best classifier in $\mathcal{H}$, and $\nu := l(h^\star)$ to be its error rate. Let γ_2 to be an absolute constant to be specified in Lemma C.14 in Appendix.

We introduce some definitions that will be used to upper-bound the size of the disagreement sets in our algorithm. Let $\text{DIS}_0 := \mathcal{X}$. Recall $K = \lceil \log n \rceil$. For $k = 1, \ldots, K$, let $\zeta_k := \sup_{x \in \text{DIS}_{k-1}} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_{k-1} Q_0(x) + n_{k-1}}$, $\varepsilon_k := \gamma_2 \zeta_k + \gamma_2 \sqrt{\zeta_k l(h^\star)}$, $\text{DIS}_k := \text{DIS}(B(h^\star, 2\nu + \varepsilon_k))$. Let $\zeta := \sup_{x \in \text{DIS}_1} \frac{1}{\alpha Q_0(x) + 1}$.

The following theorem gives statistical consistency of our algorithm.

Theorem 6.1. *There is an absolute constant c_0 such that for any $\delta > 0$, with probability at least $1 - \delta$,*

$$l(\hat{h}) \leq l(h^\star) + c_0 \sup_{x \in DIS_K} \frac{\log \frac{K|\mathcal{H}|}{\delta}}{mQ_0(x) + n} + c_0 \sqrt{\sup_{x \in DIS_K} \frac{\log \frac{K|\mathcal{H}|}{\delta}}{mQ_0(x) + n} l(h^\star)}.$$

6.5.2 Label Complexity

We first introduce the adjusted disagreement coefficient, which characterizes the rate of decrease of the query region as the candidate set shrinks.

Definition 6.2. For any measurable set $A \subseteq \mathcal{X}$, define $S(A, \alpha)$ to be

$$\bigcup_{A' \subseteq A} \left(A' \cap \left\{ x : Q_0(x) \leq \inf_{x \in A'} Q_0(x) + \frac{1}{\alpha} \right\} \right).$$

For any $r_0 \geq 2\nu$, $\alpha \geq 1$, define the adjusted disagreement coefficient $\tilde{\theta}(r_0, \alpha)$ to be

$$\sup_{r > r_0} \frac{1}{r} \mathbb{P}(S(\mathrm{DIS}(B(h^\star, r)), \alpha)).$$

The adjusted disagreement coefficient is a generalization of the standard disagreement coefficient [Han07] which has been widely used for analyzing active learning algorithms. The standard disagreement coefficient $\theta(r)$ can be written as $\theta(r) = \tilde{\theta}(r, 1)$, and clearly $\theta(r) \geq \tilde{\theta}(r, \alpha)$ for all $\alpha \geq 1$.

The following lemma is immediate from definition.

Lemma 6.3. *For any $r \geq 2\nu$, any $\alpha \geq 1$, $\mathbb{P}(S(DIS(B(h^\star, r)), \alpha)) \leq r\tilde{\theta}(r, \alpha)$.*

We can upper-bound the number of labels queried by our algorithm using the adjusted disagreement coefficient. (Recall that we only count labels queried during the online phase, and that $\alpha = 2m/3n \geq 1$)

Theorem 6.4. *There is an absolute constant c_1 such that for any $\delta > 0$, with probability at least $1 - \delta$, the number of labels queried by Algorithm 10 is at most:*

$$c_1 \tilde{\theta}(2\nu + \varepsilon_K, \alpha)(n\nu + \zeta \log n \log \frac{|\mathcal{H}| \log n}{\delta} + \log n \sqrt{n\nu\zeta \log \frac{|\mathcal{H}| \log n}{\delta}}).$$

6.5.3 Remarks

As a sanity check, note that when $Q_0(x) \equiv 1$ (i.e., all labels in the logged data are shown), our results reduce to the classical bounds for disagreement-based active learning with a warm-start.

Next, we compare the theoretical guarantees of our algorithm with some alternatives. We fix the target error rate to be $\nu+\epsilon$, assume we are given m logged data, and compare upper bounds on the number of labels required in the online phase to achieve the target error rate. Recall $\xi_0 = \inf_{x\in\mathcal{X}} Q_0(x)$. Define $\tilde{\xi}_K := \inf_{x\in\text{DIS}_K} Q_0(x)$, $\tilde{\theta} := \tilde{\theta}(2\nu, \alpha)$, $\theta := \theta(2\nu)$.

From Theorem 6.1 and 6.4 and some algebra, the number of labels required by our algorithm is $\tilde{O}\left(\nu\tilde{\theta}\cdot(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\tilde{\xi}_K)\right)$.

The first alternative is passive learning that requests all labels for $\{X_t\}_{t=m+1}^{m+n}$ and finds an empirical risk minimizer using both logged data and online data. If standard importance sampling is used, the upper bound is $\tilde{O}\left(\frac{1}{\xi_0}(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\xi_0)\right)$. If multiple importance sampling is used, the upper bound is $\tilde{O}\left(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\tilde{\xi}_K\right)$. Both bounds are worse than ours since $\nu\tilde{\theta} \leq 1$ and $\xi_0 \leq \tilde{\xi}_K \leq 1$.

A second alternative is standard disagreement-based active learning with naive warm-start where the logged data is only used to construct an initial candidate set. For standard importance sampling, the upper bound is $\tilde{O}\left(\frac{\nu\theta}{\xi_0}(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\xi_0)\right)$. For multiple importance sampling (i.e., out algorithm without the sample selection bias correction step), the upper bound is $\tilde{O}\left(\nu\theta\cdot(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\tilde{\xi}_K)\right)$. Both bounds are worse than ours since $\nu\tilde{\theta} \leq \nu\theta$ and $\xi_0 \leq \tilde{\xi}_K \leq 1$.

A third alternative is to merely use past policy to label data – that is, query on x with probability $Q_0(x)$ in the online phase. The upper bound here is $\tilde{O}\left(\frac{\mathbb{E}[Q_0(X)]}{\xi_0}(\frac{\nu+\epsilon}{\epsilon^2}\log\frac{|\mathcal{H}|}{\delta} - m\xi_0)\right)$. This is worse than ours since $\xi_0 \leq \mathbb{E}[Q_0(X)]$ and $\xi_0 \leq \tilde{\xi}_K \leq 1$.

6.6 Experiments

We now empirically validate our theoretical results by comparing our algorithm with a few alternatives on several datasets and logging policies. In particular, we confirm that the test error of our classifier drops faster than several alternatives as the expected number of label queries increases. Furthermore, we investigate the effectiveness of two key components of our algorithm: multiple importance sampling and the sample selection bias correction query strategy.

6.6.1 Methodology

Algorithms

To the best of our knowledge, no algorithms with theoretical guarantees have been proposed in the literature. We consider the overall performance of our algorithm against two natural baselines: standard passive learning (PASSIVE) and the disagreement-based active learning algorithm with warm start (DBALW). To understand the contribution of multiple importance sampling and the sample selection bias correction query strategy, we also compare the results with the disagreement-based active learning with warm start that uses multiple importance sampling (DBALWM). We do not compare with the standard disagreement-based active learning that ignores the logged data since the contribution of warm start is clear: it always results in a smaller initial candidate set, and thus leads to less label queries.

Precisely, the algorithms we implement are:

- PASSIVE: A passive learning algorithm that queries labels for all examples in the online sequence and uses the standard importance sampling estimator to combine logged data and online data.

- DBALW: A disagreement-based active learning algorithm that uses the standard importance sampling estimator, and constructs the initial candidate set with logged data. This algorithm only uses only our first key idea – warm start.

- DBALWM: A disagreement-based active learning algorithm that uses the multiple importance sampling estimator, and constructs the initial candidate set with logged data. This algorithm uses our first and second key ideas, but not the sample selection bias correction query strategy. In other words, this method sets $Q_k \equiv 1$ in Algorithm 10.

- REWEIGHTEDDBAL: The method proposed in this chapter: improved disagreement-based active learning algorithm with warm start that uses the multiple importance sampling estimator and the sample selection bias correction query strategy.

Data

Due to lack of public datasets for learning with logged data, we convert datasets for standard binary classification into our setting. Specifically, we first randomly select 80% of the whole dataset as training data and the remaining 20% is test data. We randomly select 50% of the training set as logged data, and the remaining 50% is online data. We then run an artificial logging policy (to be specified later) on the logged data to determine whether each label should be revealed to the learning algorithm or not.

Experiments are conducted on synthetic data and 11 datasets from UCI datasets [Lic13] and LIBSVM datasets [CL11]. The synthetic data is generated as follows: we generate 6000 30-dimensional points uniformly from hypercube $[-1,1]^{30}$, and labels are assigned by a random linear classifier and then flipped with probability 0.1 independently. Table 6.1 summarizes the information of datasets used in this work.

Table 6.1: Dataset information.

Dataset	# of examples	# of features
synthetic	6000	30
letter (U vs P)	1616	16
skin	245057	3
magic	19020	10
covtype	581012	54
mushrooms	8124	112
phishing	11055	68
splice	3175	60
svmguide1	4000	4
a5a	6414	123
cod-rna	59535	8
german	1000	24

We use the following four logging policies:

- IDENTICAL: Each label is revealed with probability 0.005.
- UNIFORM: We first assign each instance in the instance space to three groups with (approximately) equal probability. Then the labels in each group are revealed with probability 0.005, 0.05, and 0.5 respectively.
- UNCERTAINTY: We first train a coarse linear classifier using 10% of the data. Then, for an instance at distance r to the decision boundary, we reveal its label with probability $\exp(-cr^2)$ where c is some constant. This policy is intended to simulate uncertainty sampling used in active learning.
- CERTAINTY: We first train a coarse linear classifier using 10% of the data. Then, for an instance at distance r to the decision boundary, we reveal its label with probability cr^2 where c is some constant. This policy is intended to simulate a scenario where an action (i.e. querying for labels in our setting) is taken only if the current model is certain about its consequence.

6.6.2 Implementation

All algorithms considered in our experiments require empirical risk minimization. Instead of optimizing the 0-1 loss which is known to be computationally hard, we approximate it by optimizing a squared loss. We use the online gradient descent method in [KL11] for optimizing importance weighted loss functions.

For ReweightedDBAL, recall that in Algorithm 10, we need to find the empirical risk minimizer $\hat{h}_k \leftarrow \arg\min_{h\in C_k} l(h,\tilde{S}_k)$, update the candidate set $C_{k+1} \leftarrow \{h \in C_k \mid l(h,\tilde{S}_k) \leq l(\hat{h}_k,\tilde{S}_k) + \Delta_k(h,\hat{h}_k)\}$, and check whether $x \in \mathrm{DIS}(C_{k+1})$.

In our experiment, we approximately implement this following Vowpal Wabbit [vw]. More specifically,

1. Instead of optimizing 0-1 loss which is known to be computationally hard, we use a surrogate loss $l(y,y') = (y-y')^2$.

2. We do not explicitly maintain the candidate set C_{k+1}.

3. To solve the optimization problem $\min_{h\in C_k} l(h,\tilde{S}_k) = \sum_{(X,\tilde{Y},Z)\in\tilde{S}_k} \frac{\mathbb{1}\{h(X)\neq\tilde{Y}\}Z}{m_k Q_0(X)+n_k Q_k(X)}$, we ignore the constraint $h \in C_k$, and use online gradient descent with stepsize $\sqrt{\frac{\eta}{t+\eta}}$ where η is a parameter. The start point for gradient descent is set as $\hat{h}_{k-1}$ the ERM in the last iteration, and the step index t is shared across all iterations (i.e. we do not reset t to 1 in each iteration).

4. To approximately check whether $x \in \mathrm{DIS}(C_{k+1})$, when the hypothesis space $\mathcal{H}$ is linear classifiers, let w_k be the normal vector for current ERM $\hat{h}_k$, and a be current stepsize. We claim $x \in \mathrm{DIS}(C_{k+1})$ if $\frac{|2w_k^\top x|}{ax^\top x} \leq \sqrt{\frac{C\cdot l(\hat{h}_k,\tilde{S}_k)}{m_k\xi_k+n_k}} + \frac{C\log(m_k+n_k)}{m_k\xi_k+n_k}$ (recall $|\tilde{S}_k| = m_k+n_k$ and $\xi_k = \inf_{x\in\mathrm{DIS}(C_k)} Q_0(x)$) where C is a parameter that captures the model capacity. See [KL11] for the rationale of this approximate disagreement test.

5. $\xi_k = \inf_{x \in \text{DIS}(C_k)} Q_0(x)$ can be approximately estimated with a set of unlabeled samples. This estimate is always an upper bound of the true value of ξ_k.

DBALw and DBALwm can be implemented similarly.

Metrics and Parameter Tuning

The experiments are conducted as follows. For a fixed policy, for each dataset d, we repeat the following process 10 times. At time k, we first randomly generate a simulated logged dataset, an online dataset, and a test dataset as stated above. Then for $i = 1, 2, \cdots$, we set the horizon of the online data stream $a_i = 10 \times 2^i$ (in other words, we only allow the algorithm to use first a_i examples in the online dataset), and run algorithm A with parameter set p (to be specified later) using the logged dataset and first a_i examples in the online dataset. We record $n(d,k,i,A,p)$ to be the number of label queries, and $e(d,k,i,A,p)$ to be the test error of the learned linear classifier.

Let $\bar{n}(d,i,A,p) = \frac{1}{10}\sum_k n(d,k,i,A,p)$, $\bar{e}(d,i,A,p) = \frac{1}{10}\sum_k e(d,k,i,A,p)$. To evaluate the overall performance of algorithm A with parameter set p, we use the following area under the curve metric (see also [HAH$^+$15]):

$$\begin{aligned}\text{AUC}(d,A,p) = \sum_i & \frac{\bar{e}(d,i,A,p) + \bar{e}(d,i+1,A,p)}{2} \\ & \cdot (\bar{n}(d,i+1,A,p) - \bar{n}(d,i,A,p)).\end{aligned}$$

A small value of AUC means that the test error decays fast as the number of label queries increases.

The parameter set p consists of two parameters:

- Model capacity C. In our theoretical analysis there is a term $C := O(\log \frac{\mathcal{H}}{\delta})$ in the bounds, which is known to be loose in practice [Hsu10]. Therefore, in experiments, we treat C as a parameter to tune. We try C in $\{0.01 \times 2^k \mid k = 0, 2, 4, \dots, 18\}$

- Learning rate η. We use online gradient descent with stepsize $\sqrt{\frac{\eta}{t+\eta}}$. We try η in $\{0.0001 \times 2^k \mid k = 0, 2, 4, \dots, 18\}$.

For each policy, we report $\text{AUC}(d, A) = \min_p \text{AUC}(d, A, p)$, the AUC under the parameter set that minimizes AUC for dataset d and algorithm A.

6.6.3 Results and Discussion

We report the AUCs for each algorithm under each policy and each dataset in Tables 6.2 to 6.5, and test error curves in Figures 6.1 to 6.4.

Overall Performance The results confirm that the test error of the classifier output by our algorithm (REWEIGHTEDDBAL) drops faster than the baselines PASSIVE and DBALW: as demonstrated in Tables 6.2 to 6.5, REWEIGHTEDDBAL achieves lower AUC than both PASSIVE and DBALW for a majority of datasets under all policies. We also see that REWEIGHTEDDBAL performs better than or close to DBALWM for all policies other than Identical. This confirms that among our two key novel ideas, using multiple importance sampling consistently results in a performance gain. Using the sample selection bias correction query strategy over multiple importance sampling also leads to performance gains, but these are less consistent.

The Effectiveness of Multiple Importance Sampling As noted previously, multiple importance sampling estimators have lower variance than standard importance sampling estimators, and thus can lead to a lower label complexity. This is verified in our experiments

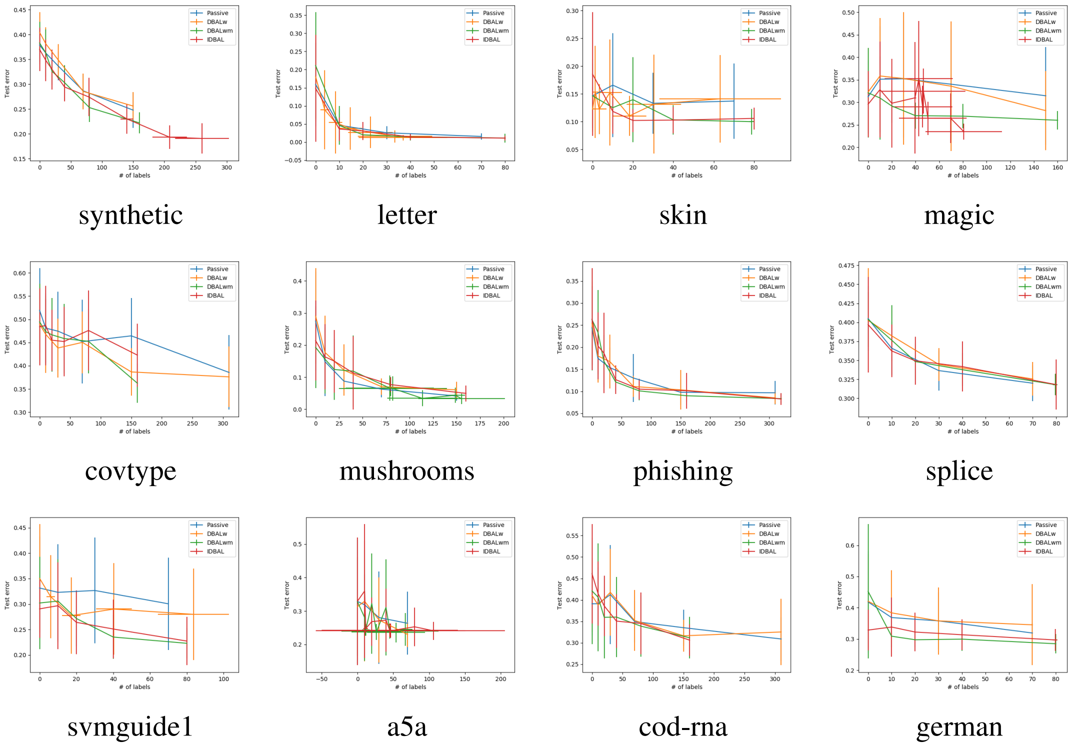
Table 6.2: AUC under Identical policy

Dataset	Passive	DBALw	DBALwm	ReweightedDBAL
synthetic	121.77	123.61	111.16	**106.66**
letter	4.40	3.65	3.82	**3.48**
skin	27.53	27.29	21.48	**21.44**
magic	109.46	101.77	89.95	**83.82**
covtype	228.04	209.56	**208.82**	220.27
mushrooms	19.22	25.29	**18.54**	23.67
phishing	78.49	73.40	**70.54**	71.68
splice	65.97	67.54	65.73	**65.66**
svmguide1	59.36	55.78	**46.79**	48.04
a5a	53.34	**50.8**	51.10	51.21
cod-rna	175.88	176.42	167.42	**164.96**
german	65.76	68.68	**59.31**	61.54

that DBALWM (DBAL with multiple importance sampling estimators) has a lower AUC than DBALW (DBAL with standard importance sampling estimator) on a majority of datasets under all policies.

The Effectiveness of the Sample Selection Bias Correction Query Strategy Under Identical policy, all labels in the logged data are revealed with equal probability. In this case, our algorithm REWEIGHTEDDBAL queries all examples in the disagreement region as DBALWM does. As shown in Table 6.2, REWEIGHTEDDBAL and DBALWM achieves the best AUC on similar number of datasets, and both methods outperform DBALW over most datasets.

Under Uniform, Uncertainty, and Certainty policies, labels in the logged data are revealed with different probabilities. In this case, REWEIGHTEDDBAL's sample selection bias correction query strategy takes effect: it queries less frequently the instances that are well-represented in the logged data, and we show that this could lead to a lower label complexity theoretically. In our experiments, as shown in Tables 6.3 to 6.5, REWEIGHTEDDBAL does indeed outperform DBALWM on these policies empirically.



synthetic

letter

skin

magic

covtype

mushrooms

phishing

splice

svmguide1

a5a

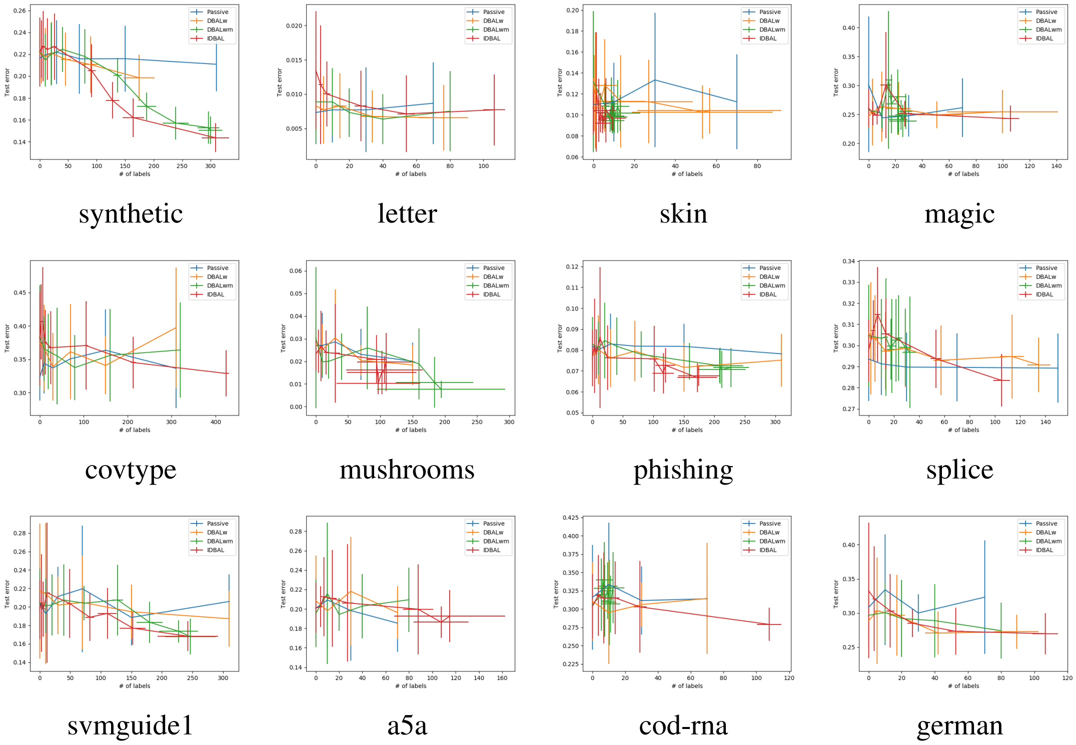
cod-rna

german

Figure 6.1: Test error vs. number of labels under the Identical policy

Table 6.3: AUC under Uniform policy

Dataset	Passive	DBALw	DBALwm	ReweightedDBAL
synthetic	113.49	106.24	92.67	**88.38**
letter	1.68	**1.29**	1.45	1.59
skin	23.76	21.42	20.67	**19.58**
magic	53.63	51.43	51.78	**50.19**
covtype	**262.34**	287.40	274.81	263.82
mushrooms	7.31	6.81	**6.51**	6.90
phishing	42.53	39.56	39.19	**37.02**
splice	88.61	89.61	90.98	**87.75**
svmguide1	110.06	105.63	98.41	**96.46**
a5a	**46.96**	48.79	49.50	47.60
cod-rna	63.39	63.30	66.32	**58.48**
german	63.60	55.87	56.22	**55.79**



synthetic letter skin magic

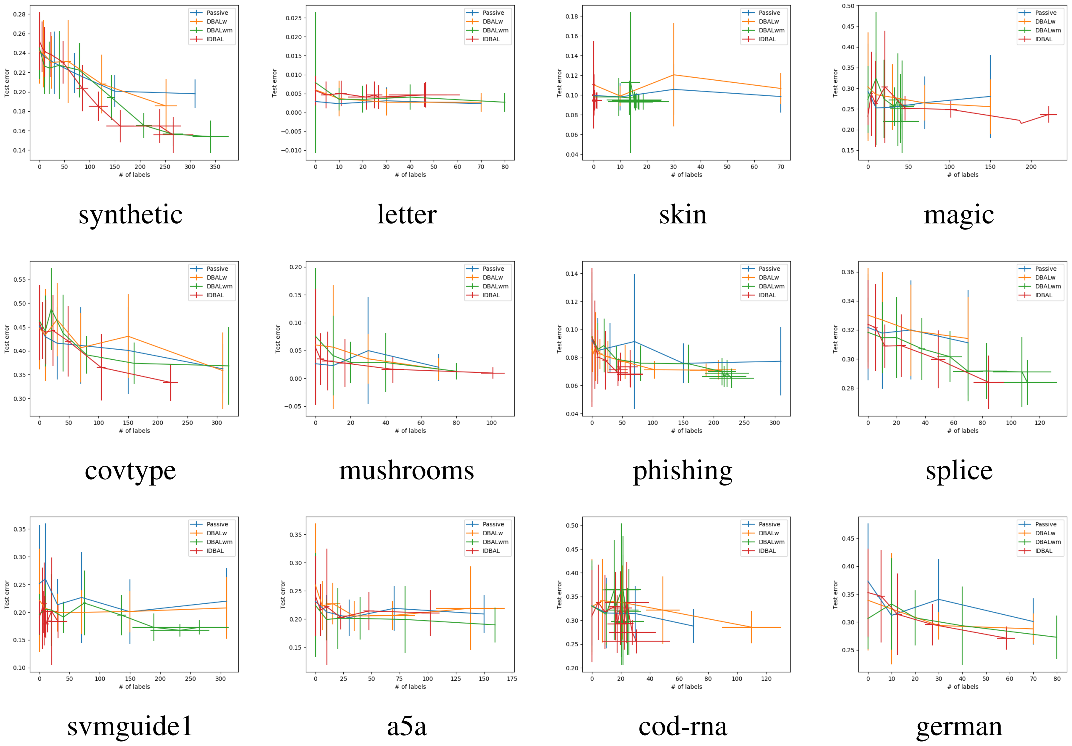
covtype mushrooms phishing splice

svmguide1 a5a cod-rna german

Figure 6.2: Test error vs. number of labels under the Uniform policy

Table 6.4: AUC under Uncertainty policy

Dataset	Passive	DBALw	DBALwm	ReweightedDBAL
synthetic	117.86	113.34	100.82	**99.1**
letter	**0.65**	0.70	0.71	1.07
skin	20.19	21.91	**18.89**	19.10
magic	106.48	101.90	99.44	**90.05**
covtype	272.48	274.53	271.37	**251.56**
mushrooms	4.93	4.64	3.77	**2.87**
phishing	52.96	48.62	**46.55**	46.59
splice	62.94	63.49	60.00	**58.56**
svmguide1	117.59	111.58	**98.88**	100.44
a5a	70.97	72.15	**65.37**	69.54
cod-rna	60.12	61.66	64.48	**53.38**
german	62.64	58.87	56.91	**56.67**



synthetic letter skin magic

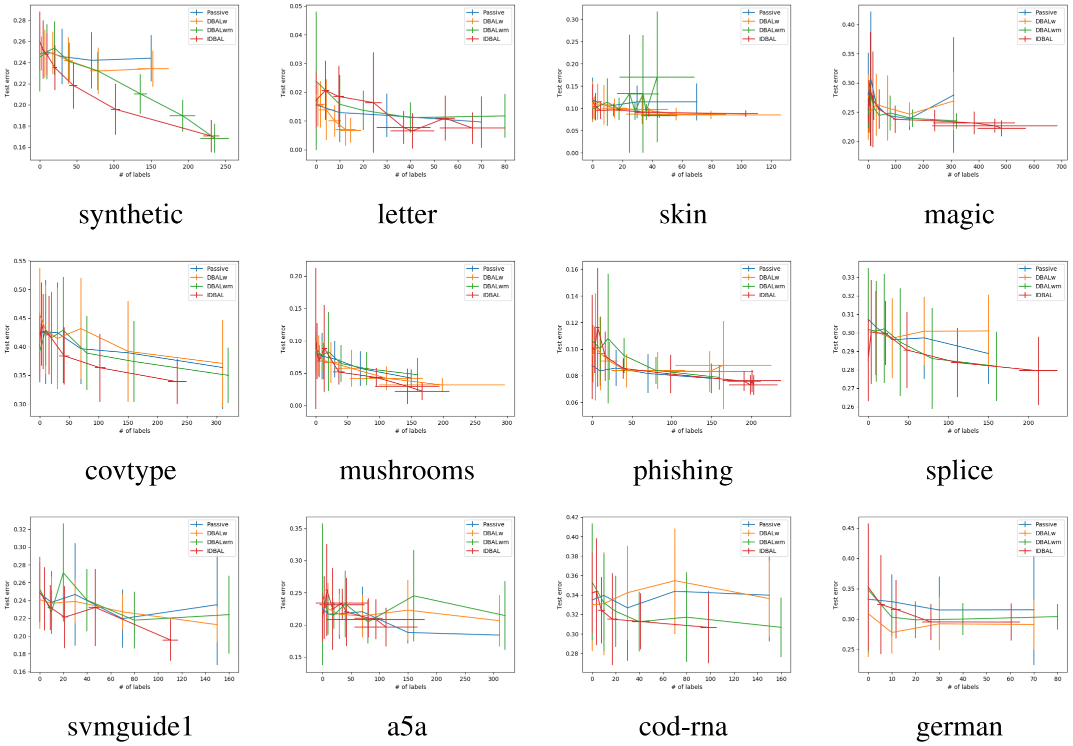
covtype mushrooms phishing splice

svmguide1 a5a cod-rna german

Figure 6.3: Test error vs. number of labels under the Uncertainty policy

Table 6.5: AUC under Certainty policy

Dataset	Passive	DBALw	DBALwm	ReweightedDBAL
synthetic	114.86	111.02	92.39	**88.82**
letter	2.02	**1.43**	2.46	1.87
skin	22.89	**17.92**	18.17	18.11
magic	231.64	225.59	205.95	**202.29**
covtype	235.68	240.86	228.94	**216.57**
mushrooms	16.53	14.62	17.97	**11.65**
phishing	34.70	37.83	35.28	**33.73**
splice	125.32	129.46	122.74	**122.26**
svmguide1	94.77	91.99	92.57	**84.86**
a5a	**119.51**	132.27	138.48	125.53
cod-rna	98.39	98.87	90.76	**90.2**
german	63.47	**58.05**	61.16	59.12



synthetic

letter

skin

magic

covtype

mushrooms

phishing

splice

svmguide1

a5a

cod-rna

german

Figure 6.4: Test error vs. number of labels under the Certainty policy

6.7 Acknowledgements

This chapter is based on the material in International Conference on Machine Learning 2018 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”Active Learning with Logged Data”) [YCJ18]. The dissertation author is the co-primary investigator and co-author of this material.

Chapter 7

Active Learning with Logged Observational Data II: A Solution with Reduced Variance

7.1 Introduction

In this chapter, we continue studying learning with logged observational data. In the previous section, we propose a modified version of disagreement-based active learning [CAL94, DHM07, BBL09, Han14], along with an importance weighted empirical risk to account for the population. However, a problem with this approach is that the importance weighted risk estimator can have extremely high variance when the importance weights – that reflect the inverse of how frequently an instance in the population is selected by the policy – are high; this may happen if, for example, certain patients are rarely given the treatment. This high variance in turn results in high label requirement for the learner.

The problem of high variance in the loss estimator is addressed in the passive case by minimizing a form of counterfactual risk [SJ15a] – an importance weighted loss that combines a variance regularizer and importance weight clipping or truncation to achieve low generalization error. A plausible solution is to use this risk for active learning as well. However, this cannot be readily achieved for two reasons. The first is that the variance regularizer itself is a function of the entire dataset, and is therefore challenging to use in interactive learning where data arrives sequentially. The second reason is that the minimizer of the (expected) counterfactual risk depends on n, the data size, which again is inconvenient for learning in an interactive manner.

In this work, we address both challenges. To address the first, we use, instead of a variance regularizer, a novel regularizer based on the second moment; the advantage is that it decomposes across multiple segments of the dataset as which makes it amenable for active learning. We provide generalization bounds for this modified counterfactual risk minimizer, and show that it has almost the same performance as counterfactual risk minimization with a variance regularizer [SJ15a]. The second challenge arises because disagreement-based active learning ensures statistical consistency by maintaining a set of plausible minimizers of the expected risk. This is problematic when the minimizer of the expected risk itself changes between iterations as in the case with our modified regularizer. We address this challenge by introducing a novel variant of disagreement-based active learning which is always guaranteed to maintain the population error minimizer in its plausible set.

Additionally, to improve sample efficiency, we then propose a third novel component – a new sampling algorithm for correcting sample selection bias that selectively queries labels of those examples which are underrepresented in the observational data. Combining these three components gives us a new algorithm. We prove this newly proposed algorithm is statistically consistent – in the sense that it converges to the true minimizer of the population risk given enough data. We also analyze its label complexity, show it is better than the algorithm we derive in Chapter 6, and demonstrate the contribution of each component of the algorithm to the label complexity bound.

7.2 Variance-Controlled Importance Sampling

In the passive setting, the standard method to overcome sample selection bias is to optimize the importance weighted (IW) loss $l(h,T_0) = \frac{1}{m}\sum_t \frac{\mathbb{1}\{h(X_t)\neq Y_t\}Z_t}{Q_0(X_t)}$. This loss is an unbiased estimator of the population error $\mathbb{P}(h(X)\neq Y)$, but its variance $\frac{1}{m}\mathbb{E}(\frac{\mathbb{1}\{h(X)\neq Y\}Z}{Q_0(X)} - l(h))^2$ can be high, leading to poor solutions. Previous work addresses this issue by adding a variance regularizer [MP09, SJ15a, ND17] and clipping/truncating the importance weight [BPQC$^+$13, SJ15a]. However, the variance regularizer is challenging to use in interactive learning when data arrives sequentially, and it is unclear how the clipping/truncating threshold should be chosen to yield good theoretical guarantees.

In this chapter, as an alternative to the variance regularizer, we propose a novel second moment regularizer which achieves a similar error bound to the variance regularizer [ND17]; and this motivates a principled choice of the clipping threshold.

7.2.1 Second-Moment-Regularized Empirical Risk Minimization

Intuitively, between two classifiers with similarly small training loss $l(h,T_0)$, the one with lower variance should be preferred, since its population error $l(h)$ would be small with a higher probability than the one with higher variance. Existing work encourages low variance by regularizing the loss with the estimated variance $\hat{\mathrm{Var}}(h,T_0) = \frac{1}{m}\sum_i(\frac{\mathbb{1}\{h(X_i)\neq Y_i\}Z_i}{Q_0(X_i)})^2 - l(h,T_0)^2$. Here, we propose to regularize with the estimated second moment $\hat{\mathrm{V}}(h,T_0) = \frac{1}{m}\sum_i(\frac{\mathbb{1}\{h(X_i)\neq Y_i\}Z_i}{Q_0(X_i)})^2$, an upper bound of $\hat{\mathrm{Var}}(h,T_0)$. We have the following generalization error bound for regularized ERM.

Theorem 7.1. *Let* $\hat{h} = \arg\min_{h\in\mathcal{H}} l(h,T_0) + \sqrt{\frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\hat{V}(h,T_0)}$*. For any* $\delta > 0$*, then with probability at least* $1-\delta$*,* $l(\hat{h}) - l(h^\star) \le \frac{28\log\frac{|\mathcal{H}|}{\delta}}{3mq_0} + \sqrt{\frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{Q_0(X)}} + \frac{\sqrt{4\log\frac{|\mathcal{H}|}{\delta}}}{m^{\frac{3}{2}}q_0^2}$*.*

Theorem 7.1 shows a error rates similar to the one for the variance regularizer [ND17]. However, the advantage of using the second moment is the decomposability: $\hat{V}(h,S_1\cup S_2) = \frac{|S_1|}{|S_1|+|S_2|}\hat{V}(h,S_1) + \frac{|S_2|}{|S_1|+|S_2|}\hat{V}(h,S_2)$. This makes it easier to analyze for active learning that we will discuss later.

Recall for $\hat{h}_{\mathrm{IW}} = \arg\min_{h\in\mathcal{H}} l(h,T_0)$, the unregularized importance sampling loss minimizer , the error bound is $\tilde{O}(\frac{\log|\mathcal{H}|}{mq_0} + \sqrt{\frac{\log|\mathcal{H}|}{m}\min(\frac{l(h^\star)}{q_0},\mathbb{E}\frac{1}{Q_0(X)})})$ [CMM10, YCJ18]. In Theorem 7.1, the extra $\frac{1}{m^{\frac{3}{2}}q_0^2}$ term is due to the deviation of $\sqrt{\hat{V}(h,T_0)}$ around $\sqrt{\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{Q_0(X)}}$, and is negligible when m is large. In this case, learning with a second moment regularizer gives a better generalization bound.

This improvement in generalization error is due to the regularizer instead of tighter analysis. Similar to [MP09, ND17], we show in Theorem 7.2 that for some distributions, the error bound in Theorem 7.1 cannot be achieved by any algorithm that simply optimizes the unregularized empirical loss.

Theorem 7.2. *For any* $0 < \nu < \frac{1}{3}$*,* $m \ge \frac{49}{\nu^2}$*, there is a sample space* $\mathcal{X}\times\mathcal{Y}$*, a hypothesis class* $\mathcal{H}$*, a distribution* D*, and a logging policy* Q_0 *such that* $\frac{\nu}{q_0} > \mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{Q_0(X)}$*, and that with probability at least* $\frac{1}{100}$ *over the draw of* $S = \{(X_t,Y_t,Z_t)\}_{t=1}^m$*, if* $\hat{h} = \arg\min_{h\in\mathcal{H}} l(h,S)$*, then* $l(\hat{h}) \ge l(h^\star) + \frac{1}{mq_0} + \sqrt{\frac{\nu}{mq_0}}$*.*

Proof. (of Theorem 7.2) For any $0 < \nu < \frac{1}{3}$, $m > \frac{49}{\nu^2}$, set $q_0 = \frac{1}{40}\nu$, $c = \frac{1}{3}$, $\varepsilon = \frac{c^2+\sqrt{c^4+4c^2q_0\nu m}}{2q_0m}$. It can be checked that $\varepsilon < \nu$ and $m = c^2\frac{\nu+\varepsilon}{q_0\varepsilon^2}$. Let $\mathcal{X} = \{x_1,x_2,x_3\}$, and define $\mathbb{P}(X = x_1) = \nu$,

$\mathbb{P}(X = x_2) = \nu + \varepsilon$, $\mathbb{P}(X = x_3) = 1 - 2\nu - \varepsilon$, and $\mathbb{P}(Y = 1) = 1$. Let $\mathcal{H} = \{h_1, h_2\}$ where $h_1(x_1) = -1$, $h_1(x_2) = h_1(x_3) = 1$, and $h_2(x_2) = -1$, $h_2(x_1) = h_2(x_3) = 1$. Define the logging policy $Q_0(x_1) = Q_0(x_3) = 1$, $Q_0(x_2) = q_0$. Let $S = \{(X_t, Y_t, Z_t)\}_{t=1}^m$ be a dataset of size m generated from the aforementioned distribution. Clearly, we have $l(h_1) = \nu$ and $l(h_2) = \nu + \varepsilon$. We next prove that $\mathbb{P}(l(h_1,S) > l(h_2,S)) \geq \frac{1}{100}$. This implies that with probability at least $\frac{1}{100}$,h_2 is the minimizer of the importance weighted loss $l(h,S)$, and its population error $\mathbb{P}(h_2(X) \neq Y) = \nu + \varepsilon = \nu + \frac{1}{q_0 m} + \sqrt{\frac{\nu}{q_0 m}}$.

We have

$$\begin{aligned}
\mathbb{P}(l(h_1,S) > l(h_2,S)) &\geq \mathbb{P}(l(h_1,S) > \nu - \frac{\varepsilon}{2} \text{ and } l(h_2,S) < \nu - \frac{\varepsilon}{2}) \\
&= 1 - \mathbb{P}(l(h_1,S) \leq \nu - \frac{\varepsilon}{2} \text{ or } l(h_2,S) \geq \nu - \frac{\varepsilon}{2}) \\
&\geq 1 - \mathbb{P}(l(h_1,S) \leq \nu - \frac{\varepsilon}{2}) - \mathbb{P}(l(h_2,S) \geq \nu - \frac{\varepsilon}{2}) \\
&= \mathbb{P}(l(h_2,S) < \nu - \frac{\varepsilon}{2}) - \mathbb{P}(l(h_1,S) \leq \nu - \frac{\varepsilon}{2})
\end{aligned}$$

Observe that by our construction, $ml(h_1,S) = \sum_{i=1}^m \mathbb{1}\{X_i = x_1\}$ follows the binomial distribution $\text{Bin}(m,\nu)$. By a Chernoff bound, $\mathbb{P}(l(h_1,S) \leq \nu - \frac{\varepsilon}{2}) \leq e^{-\frac{1}{2}m\varepsilon^2}$. Since $\varepsilon \geq \sqrt{\frac{c^2\nu}{q_0 m}} \geq \sqrt{\frac{40c^2}{m}}$, $e^{-\frac{1}{2}m\varepsilon^2} \leq e^{-20c^2} = e^{-\frac{20}{9}}$.

By our construction, we also have that $q_0 ml(h_2,S) = \sum_{i=1}^m \mathbb{1}\{X_i = x_2, Z_i = 1\}$ which follows the binomial distribution $\text{Bin}(m, q_0(\nu+\varepsilon))$. Thus, $\mathbb{P}(l(h_2,S) \leq \nu - \frac{\varepsilon}{2}) = \mathbb{P}(q_0 ml(h_2,S) \leq q_0 m(\nu+\varepsilon) - \frac{3}{2}q_0 m\varepsilon) \geq \frac{1}{\sqrt{2\pi}}\frac{3c}{9c^2+1}e^{-\frac{9}{2}c^2} = \frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}}$ where the inequality follows by Lemma D.9.

Therefore, $\mathbb{P}(l(h_1,S) > l(h_2,S)) \geq \mathbb{P}(l(h_2,S) < \nu - \frac{\varepsilon}{2}) - \mathbb{P}(l(h_1,S) \leq \nu - \frac{\varepsilon}{2}) \geq \frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}} - e^{-\frac{20}{9}} \geq \frac{1}{100}$. □

Remark 7.3. A similar result for general cost-sensitive empirical risk minimization is proved in [MP09, ND17]. In [MP09, ND17], they construct examples where $\mathrm{Var}(h^\star) = 0$ and learning $h^\star$ with unregularized ERM gives $\tilde{\Omega}(\sqrt{\frac{1}{m}})$ error, while regularized ERM gives $\tilde{O}(\frac{1}{m})$ error. However, their construction does not work in our setting because the bound for unregularized ERM (Chapter 6) also gives $\tilde{O}(\frac{1}{m})$ error when $\mathrm{Var}(h^\star) = 0$ (since $\mathrm{Var}(h^\star) = 0$ implies $l(h^\star) = 0$), so more careful construction and analysis are needed.

7.2.2 Clipped Importance Sampling

The variance and hence the error bound for second-moment regularized ERM can still be high if $\frac{1}{Q_0(x)}$ is large. This $\frac{1}{Q_0(X)}$ factor arises inevitably to guarantee the importance weighted estimator is unbiased. Existing work alleviates the variance issue at the cost of some bias by clipping or truncating the importance weight. In this chapter, we focus on clipping, where the loss estimator becomes $l(h; T_0, M) := \frac{1}{m}\sum_{i=1}^{m} \frac{\mathbb{1}\{h(X_i) \neq Y_i\} Z_i}{Q_0(X_i)} \mathbb{1}[\frac{1}{Q_0(X_i)} \leq M]$. This estimator is no longer unbiased, but as the weight is clipped at M, so is the variance. Although studied previously [BPQC$^+$13, SJ15a], to the best of our knowledge, it remains unclear how the clipping threshold M can be chosen in a principled way.

We propose to choose $M_0 = \inf\{M' \geq 1 \mid \frac{2M' \log\frac{|\mathcal{H}|}{\delta}}{m} \geq \mathbb{P}_X(\frac{1}{Q_0(X)} > M')\}$. This choice of M_0 is chosen to minimize the following error bound for the clipped second-moment regularized ERM (proved in Theorem D.20 in Appendix):

$$\begin{aligned} l(\hat{h}_M) - l(h^\star) \leq & \frac{2\lambda M}{m} + \frac{16M}{3m}\log\frac{|\mathcal{H}|}{\delta} + \frac{M^2}{m^{\frac{3}{2}}}\sqrt{4\log\frac{|\mathcal{H}|}{\delta}} \\ & + \sqrt{\frac{\lambda}{m}\mathbb{E}\frac{\mathbb{1}\{h^\star(X) \neq Y\}}{Q_0(X)}\mathbb{1}[\frac{1}{Q_0(X)} \leq M]} + \mathbb{P}_X(\frac{1}{Q_0(X)} > M). \end{aligned}$$

In particular, to choose M that minimizes the RHS,we set $\lambda = 4\log\frac{|\mathcal{H}|}{\delta}$, focus on the low order terms with respect to m, and minimize $e(M) := \sqrt{\frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\mathbb{E}\frac{1}{Q_0(X)}\mathbb{1}[\frac{1}{Q_0(X)} \leq M] + \mathbb{P}_X(\frac{1}{Q_0(X)} >}$ $M)$ instead since $\mathbb{1}\{h^\star(X) \neq Y\}$ could not be determined with unlabeled samples. In this sense, the following proposition shows that our choice of M is nearly optimal.

Proposition 7.4. *Suppose random variable $\frac{1}{Q_0(X)}$ has a probability density function, and there exists $M_0 \geq 1$ such that $\frac{2\log\frac{|\mathcal{H}|}{\delta}}{m}M_0 = \mathbb{P}_X(\frac{1}{Q_0(X)} > M_0)$. Then $e(M_0) \leq \sqrt{2}\inf_{M\geq 1} e(M)$.*

Proof. Define $f_1(M) = \frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\mathbb{E}\frac{1}{Q_0(X)}\mathbb{1}[\frac{1}{Q_0(X)} \leq M]$, and $f_2(M) = \mathbb{P}_X(\frac{1}{Q_0(X)} > c)$. We first show that $f_1(M_0) + f_2(M_0)^2 \leq \inf_{M>1} f_1(M) + f_2(M)^2$.

Let $g(x)$ be the probability density function of random variable $1/Q_0(X)$. We have $f_1(M) = \frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\int_0^M xg(x)\,\mathrm{d}x$ and $f_2(M) = \int_M^\infty g(x)\,\mathrm{d}x$, so $f_1'(M) = \frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}Mg(M)$, and $f_2'(M) = -g(M)$. Define $f(M) = f_1(M) + f_2(M)^2$. We have

$$\begin{aligned} f'(M) &= f_1'(M) + 2f_2'(M)f_2(M) \\ &= 2g(M)(\frac{2\log\frac{|\mathcal{H}|}{\delta}}{m}M - f_2(M)). \end{aligned}$$

Recall we assume there exists $M_0 \geq 1$ such that $\frac{2\log\frac{|\mathcal{H}|}{\delta}}{m}M_0 = f_2(M_0)$. Since $\frac{2\log\frac{|\mathcal{H}|}{\delta}}{m}M$ is strictly increasing w.r.t. M and $f_2(M)$ is non-increasing w.r.t. M, it follows that $f(M)$ achieves its minimum at M_0, that is, for any $c \geq 1$, $f_1(M_0) + f_2^2(M_0) \leq f_1(M) + f_2^2(M)$.

Now, $\sqrt{f_1(M_0) + f_2^2(M_0)} \geq \frac{1}{\sqrt{2}}(\sqrt{f_1(M_0)} + f_2(M_0))$ since $\sqrt{a+b} \geq \frac{1}{\sqrt{2}}(\sqrt{a} + \sqrt{b})$ for any $a, b \geq 0$, and $\sqrt{f_1(M) + f_2^2(M)} \leq \sqrt{f_1(M)} + f_2(M)$ since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Thus $\frac{1}{\sqrt{2}}(\sqrt{f_1(M_0)} + f_2(M_0)) \leq \sqrt{f_1(M)} + f_2(M)$ for all $M > 0$, which concludes the proof. □

Remark 7.5. Since $\frac{1}{M}\mathbb{P}_X(\frac{1}{Q_0(X)} > M)$ is monotonically decreasing with respect to M and its range is $(0, 1)$, the existence and uniqueness of M_0 are guaranteed if $\frac{2}{m}\log\frac{|\mathcal{H}|}{\delta} < 1$.

The choice of M_0 implies that the clipping threshold should be larger as the sample size m increases, which confirms the intuition that with a larger sample size the variance becomes less of an issue than the bias. We have the following generalization error bound.

Theorem 7.6. *Let* $\hat{h} = \arg\min_{h\in\mathcal{H}} l(h;T_0,M_0) + \sqrt{\frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\hat{V}(h;T_0,M_0)}$. *For any* $\delta > 0$, *with probability at least* $1-\delta$,

$$l(\hat{h}) - l(h^\star) \leq \frac{34\log\frac{|\mathcal{H}|}{\delta}}{3m}M_0 + \frac{\sqrt{4\log\frac{|\mathcal{H}|}{\delta}}}{m^{\frac{3}{2}}}M_0^2 + \sqrt{\frac{4\log\frac{|\mathcal{H}|}{\delta}}{m}\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{Q_0(X)}\mathbb{1}[\frac{1}{Q_0(X)} \leq M_0]}.$$

We always have $M_0 \leq \frac{1}{q_0}$ as $\mathbb{P}_X(\frac{1}{Q_0(X)} > \frac{1}{q_0}) = 0$. Thus, this error bound is always no worse than that without clipping asymptotically.

The following example shows that our choice of M indeed avoids outputting suboptimal classifiers.

Example 7.7. Let $\mathcal{X} = \{x_0,x_1,x_2,x_3,x_4\}$, $\mathcal{H} = \{h_1,h_2,h_3,h_4\}$. Suppose $\mathbb{P}(Y=1) = -1$, $\nu < \frac{1}{10}$, $\alpha < 0.01$, and $\varepsilon = \frac{\nu}{1+1/100\alpha}$. The marginal distribution on X, the prediction of each classifier, and the logging policy Q_0 is defined in Table 7.1.

Table 7.1: An example for clipping

	x_0	x_1	x_2	x_3	x_4
$h_1(\cdot)$	1	1	-1	-1	-1
$h_2(\cdot)$	1	-1	1	-1	-1
$h_3(\cdot)$	1	-1	-1	1	-1
$h_4(\cdot)$	-1	-1	-1	-1	1
$\mathbb{P}_X(\cdot)$	$\nu-\varepsilon$	ε	4ε	16ε	$1-\nu-20\varepsilon$
$Q_0(\cdot)$	1	α	α	4α	4α

We have $l(h_1) = \nu$, $l(h_2) = \nu+3\varepsilon$, $l(h_3) = \nu+15\varepsilon$, $l(h_4) = 1-\nu-20\varepsilon$. Next, we consider when examples with Q_0 equals α, i.e. examples on x_1 and x_2, should be clipped. We set the failure probability $\delta = 0.01$.

If $m \geq \frac{28}{\alpha\varepsilon}$, without clipping our error bound guarantees that (by minimizing a regularized training error) learner can achieve an error of less than $\nu + 3\varepsilon$, so it would output the optimal classifier h_1 with high probability. On the other hand, if $M < \frac{1}{\alpha}$, then all examples on x_1 and x_2 are ignored due to clipping, so the learner would not be able to distinguish between $h_1$and h_2, and thus with constant probability the error of the output classifier is at least $l(h_2) = \nu + 3\varepsilon$. This means if $m \geq \frac{28}{\alpha\varepsilon}$, examples on x_1 and x_2 should not be clipped.

If $m \geq \frac{2}{\alpha\varepsilon}$ and examples on x_1 and x_2 are clipped, our error bound guarantees learner can achieve an error of less than $\nu + 16\varepsilon$, which means the learner would output either h_1 or h_2 and achieve an actual error of at most $\nu + 3\varepsilon$. However, without clipping, the learner would require $m \geq \frac{4}{\alpha\varepsilon}$ to achieve an error of less than $\nu + 16\varepsilon$. Thus, if $m \leq \frac{4}{\alpha\varepsilon}$, examples on x_1 and x_2 should be clipped.

To sum up, examples with Q_0 equals α (i.e. x_1 and x_2) should be clipped if $m \leq \frac{4}{\alpha\varepsilon}$ and not be clipped if $m \geq \frac{28}{\alpha\varepsilon}$. Our choice of the clipping threshold clips x_1 and x_2 whenever $m \leq \frac{24}{5\alpha\varepsilon}$, which falls inside the desired interval.

7.3 Active Learning

Next, we consider active learning where in addition to a logged observational dataset the learner has access to a stream of unlabeled samples from which it can actively query for labels. The main challenges are how to control the variance due to the observational data with active learning, and how to leverage the logged observational data to reduce the number of label queries beyond simply using them for warm-start.

To address these challenges, we first propose a nontrivial change to the Disagreement-Based Active Learning (DBAL) so that the variance-controlled importance sampling objective can be incorporated. This modified algorithm also works in a general cost-sensitive active learning setting which we believe is of independent interest. Second, we show how to combine logged observational data with active learning through multiple importance sampling (MIS). Finally, we propose a novel sample selection bias correction technique to query regions under-explored in the observational data more frequently. We provide theoretical analysis demonstrating that the proposed method gives better label complexity guarantees than previous work (Chapter 6) and other alternative methods.

Key Technique 1: Disagreement-Based Active Learning with Variance-Controlled Importance Sampling

The DBAL framework, presented in Algorithm 1, is a widely-used general framework for active learning The classical DBAL framework only considers the unregularized 0-1 loss. As discussed in the previous section, with observational data, unregularized loss leads to suboptimal label complexity. However, directly adding a regularizer breaks the statistical consistency of DBAL, since the proof of its consistency is contingent on two properties: (1) the minimizer of the population loss $l(h)$ stays in all candidate sets with high probability; (2) the loss difference $l(h_1,S) - l(h_2,S)$ for any $h_1, h_2 \in C_t$ does not change no matter how examples outside the disagreement region D_t are labeled.

Unfortunately, if we add a variance based regularizer (either estimated variance or second moment), the objective function $l(h,S) + \sqrt{\frac{\lambda}{n}\hat{\mathrm{V}}(h,S)}$ has to change as the sample size n increases, and so does the optimal classifier w.r.t. regularized population loss $\tilde{h}_n = \arg\min l(h) + \sqrt{\frac{\lambda}{n}V(h)}$. Consequently, $\tilde{h}_n$ may not stay in all candidate sets. Besides, the difference of the regularized loss $l(h_1,S)+\sqrt{\frac{\lambda}{n}\hat{\mathrm{V}}(h_1,S)}-(l(h_2,S)+\sqrt{\frac{\lambda}{n}\hat{\mathrm{V}}(h_2,S)})$ changes if labels of examples outside the disagreement region D_t are modified, breaking the second property.

To resolve the consistency issues, we first carefully choose the definition of the candidate set and guarantee the optimal classifier w.r.t. the prediction error $h^\star = \arg\min l(h)$, instead of the regularized loss $\tilde{h}_n$, stays in candidate sets with high probability. Moreover, instead of the plain variance regularizer, we apply the second moment regularizer and exploit its decomposability property to bound the difference of the regularized loss for ensuring consistency.

Key Technique 2: Multiple Importance Sampling

MIS addresses how to combine logged observational data with actively collected data for training classifiers [ABSJ17, YCJ18]. To illustrate this, for simplicity, we assume a fixed query policy Q_1 is used for active learning. To make use of both $T_0 = \{(X_i,Y_i,Z_i)\}_{i=1}^m$ collected by Q_0 and $T_1 = \{(X_i,Y_i,Z_i)\}_{i=m+1}^{m+n}$ collected by Q_1, one could optimize the unbiased importance weighted error estimator $l_{\mathrm{IS}}(h,T_0\cup T_1) = \sum_{i=1}^m \frac{\mathbb{1}\{h(X_i)\neq Y_i\}Z_i}{(m+n)Q_0(X_i)} + \sum_{i=m+1}^{m+n} \frac{\mathbb{1}\{h(X_i)\neq Y_i\}Z_i}{(m+n)Q_1(X_i)}$ which can have high variance and lead to poor generalization error. Here, we apply the MIS estimator $l_{\mathrm{MIS}}(h,T_0\cup T_1) := \sum_{i=1}^{m+n} \frac{\mathbb{1}\{h(X_i)\neq Y_i\}Z_i}{mQ_0(X_i)+nQ_1(X_i)}$ which effectively treats the data $T_0\cup T_1$ as drawn from a mixture policy $\frac{mQ_0+nQ_1}{m+n}$. l_{MIS} is also unbiased, but has lower variance than l_{IS} and thus gives better error bounds.

Key Technique 3: Active Sample Selection Bias Correction

Another advantage to consider active learning is that the learner can apply a strategy to correct the sample selection bias, which improves label efficiency further. This strategy is inspired from the following intuition: due to sample selection bias caused by the logging policy, labels for some regions of the sample space may be less likely to be observed in the logged data, thus increasing the uncertainty in these regions. To counter this effect, during active learning, the learner should query more labels from such regions.

We formalize this intuition as follows. Suppose we would like to design a single query strategy $Q_1 : \mathcal{X} \to [0,1]$ that determines the probability of querying the label for an instance during the active learning phase. For any Q_1, we have the following generalization error bound for learning with n logged examples and m unlabeled examples from which the learner can select and query for labels (for simplicity of illustration, we use the unclipped estimator here)

$$l(h_1) - l(h_2) \leq l(h_1, S) - l(h_2, S) + \frac{4 \log \frac{2|\mathcal{H}|}{\delta}}{3(mq_0 + n)} + \sqrt{4\mathbb{E}\frac{\mathbb{1}\{h_1(X) \neq h_2(X)\}}{mQ_0(X) + nQ_1(X)} \log \frac{2|\mathcal{H}|}{\delta}}.$$

We propose to set $Q_1(x) = \mathbb{1}\{mQ_0(x) < \frac{m}{2}Q_0(x) + n\}$ which only queries instances if $Q_0(x)$ is small. This leads to fewer queries while guarantees an error bound close to the one achieved by setting $Q_1(x) \equiv 1$ that queries every instance. Example 7.8 shows the sample selection bias correction strategy indeed improves label complexity.

Example 7.8. Let $\lambda > 1$ be any constant. Suppose $\mathcal{X} = \{x_1, x_2\}$, $Q_0(x_1) = 1$, $Q_0(x_2) = \alpha$, $\mathbb{P}(x_1) = 1 - \mu$, $\mathbb{P}(x_2) = \mu$ and assume $\mu \leq \frac{1}{4\lambda}$ and $\alpha \leq \frac{\mu^2}{2\lambda}$. Assume the logged data size m is greater than twice as the online stream size n. Without the sample selection bias correction strategy, after seeing n examples, the learner queries all n examples and achieves an error bound of $\frac{4 \log \frac{2|\mathcal{H}|}{\delta}}{3(m\alpha + n)} + \sqrt{4(\frac{c\mu}{m+n} + \frac{\mu}{m\alpha + n}) \log \frac{2|\mathcal{H}|}{\delta}}$ by minimizing the regularized MIS loss. With the sample selection bias correction strategy, the learner only queries x_2, so after seeing n examples, it queries only μn examples in expectation and achieves an error bound of $\frac{4 \log \frac{2|\mathcal{H}|}{\delta}}{3(m\alpha + n)} + \sqrt{4(\frac{c\mu}{m} + \frac{\mu}{m\alpha + n}) \log \frac{2|\mathcal{H}|}{\delta}}$. With some algebra, it can be shown that to achieve the same error bound, if $\frac{\lambda\alpha}{\mu}m \leq n \leq \frac{\mu}{2}m$, then the number of queries requested by the learner without the sample selection bias correction correction strategy is at least λ times more than the number of queries for the learner with the bias correction strategy. Since this holds for any $\lambda \geq 1$, the decrease of the number of label queries due to our sample selection bias correction strategy can be significant.

The sample selection bias correction strategy is complementary to the DBAL technique. We note that a similar query strategy is proposed in Chapter 6, but the strategy here stems from a tighter analysis and can be applied with variance control techniques discussed in Section 7.2, and thus gives better label complexity guarantees as to be discussed in the analysis section.

7.3.1 Algorithm

Putting things together, our proposed algorithm is shown as Algorithm 11. It takes the logged data and an epoch schedule as input. It assumes the logging policy Q_0 and its distribution $f(x) = \mathbb{P}(Q_0(X) \leq x)$ are known (otherwise, these quantities can be estimated with unlabeled data).

Algorithm 11 uses the DBAL framework that recursively shrinks a candidate set C and its corresponding disagreement region D to save label queries by not querying examples outside D. In particular, at iteration k, it computes a clipping threshold M_k (step 5) and MIS weights $w_k(x) := \frac{m+n_k}{mQ_0(X_i)+\sum_{j=1}^{k}\tau_i Q_i(X_i)}$ which are used to define the clipped MIS error estimator and two second moment estimators

$$
\begin{aligned}
l(h;\tilde{S}_k,M_k) &:= \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k(X_i)Z_i\mathbb{1}\{h(X_i)\neq\tilde{Y}_i\}\mathbb{1}\{w_k(X_i)\leq M_k\},\\
\hat{\mathrm{V}}(h_1,h_2;\tilde{S}_k,M_k) &:= \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k^2(X_i)Z_i\mathbb{1}\{h_1(X_i)\neq h_2(X_i)\}\mathbb{1}\{w_k(X_i)\leq M_k\},\\
\hat{\mathrm{V}}(h;\tilde{S}_k,M_k) &:= \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k^2(X_i)Z_i\mathbb{1}\{h(X_i)\neq\tilde{Y}_i\}\mathbb{1}\{w_k(X_i)\leq M_k\}.
\end{aligned}
$$

The algorithm shrinks the candidate set C_{k+1} by eliminating classifiers whose estimated error is larger than a threshold that takes the minimum empirical error and the second moment into account (step 7), and defines a corresponding disagreement region $D_{k+1} = \text{DIS}(C_{k+1})$ as the set of all instances on which there are two classifiers in the candidate set C_{k+1} that predict labels differently. It derives a query policy Q_{k+1} with the sample selection bias correction strategy (step 9). At the end of iteration k, it draws τ_{k+1} unlabeled examples. For each example X with $Q_{k+1}(X) > 0$, if $X \in D_{k+1}$, the algorithm queries for the actual label Y and sets $\tilde{Y} = Y$, otherwise it infers the label and sets $\tilde{Y} = \hat{h}_k(X)$. These examples $\{X\}$ and their inferred or queried labels $\{\tilde{Y}\}$ are then used in subsequent iterations. In the last step of the algorithm, a classifier that minimizes the clipped MIS error with the second moment regularizer over all received data is returned.

Algorithm 11 Disagreement-Based Active Learning with Logged Observational Data

```
1: Input: confidence $\delta$, logged data $T_0$, epoch schedule $\tau_1, \ldots, \tau_K$, $n = \sum_{i=1}^{K} \tau_i$.
2: $\tilde{S}_0 \leftarrow T_0$; $C_0 \leftarrow \mathcal{H}$; $D_0 \leftarrow \mathcal{X}$; $n_0 = 0$
3: for $k = 0, \ldots, K-1$ do
4:     $\sigma_1(k,\delta,M) \leftarrow (\frac{M}{m+n_k} + \frac{M^2}{(m+n_k)^{\frac{3}{2}}}) \log \frac{|\mathcal{H}|}{\delta}; \sigma_2(k,\delta) = \frac{1}{m+n_k} \log \frac{|\mathcal{H}|}{\delta}; \delta_k \leftarrow \frac{\delta}{2(k+1)(k+2)}$
5:     Choose $M_k = \inf\{M \geq 1 \mid \frac{2M}{m+n_k} \log \frac{|\mathcal{H}|}{\delta_k} \geq \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} > M/2)\}$
6:     $\hat{h}_k \leftarrow \arg\min_{h \in C_k} l(h; \tilde{S}_k, M_k)$
7:     Define the candidate set $C_{k+1} \leftarrow \{h \in C_k \mid l(h;\tilde{S}_k,M_k) \leq l(\hat{h}_k;\tilde{S}_k,M_k) + \gamma_1 \sigma_1(k,\delta_k,M_k) +$
       $\gamma_1 \sqrt{\sigma_2(k,\delta_k) \hat{\mathrm{V}}(h,\hat{h}_k;\tilde{S}_k,M_k)}\}$
8:     Define the Disagreement Region $D_{k+1} \leftarrow \{x \in \mathcal{X} \mid \exists h_1, h_2 \in C_{k+1}$ s.t. $h_1(x) \neq h_2(x)\}$
9:     $Q_{k+1}(x) \leftarrow \mathbb{1}\{mQ_0(x) + \sum_{i=1}^{k} \tau_i Q_i(x) < \frac{m}{2} Q_0(x) + n_{k+1}\}$;
10:    $n_{k+1} \leftarrow n_k + \tau_{k+1}$
11:    Draw $\tau_{k+1}$ samples $\{(X_t,Y_t)\}_{t=m+n_k+1}^{m+n_{k+1}}$, and present $\{X_t\}_{t=m+n_k+1}^{m+n_{k+1}}$ to the learner.
12:    for $t = m + n_k + 1$ to $m + n_{k+1}$ do
13:        $Z_t \leftarrow Q_{k+1}(X_t)$
14:        if $Z_t = 1$ then
15:            If $X_t \in D_{k+1}$, query for label: $\tilde{Y}_t \leftarrow Y_t$; otherwise infer $\tilde{Y}_t \leftarrow \hat{h}_k(X_t)$.
16:        end if
17:    end for
18:    $\tilde{T}_{k+1} \leftarrow \{X_t, \tilde{Y}_t, Z_t\}_{t=m+n_k+1}^{m+n_{k+1}}$, $\tilde{S}_{k+1} \leftarrow \tilde{S}_k \cup \tilde{T}_{k+1}$;
19: end for
20: Output $\hat{h} = \arg\min_{h \in C_K} l(h;\tilde{S}_K,M_k) + \gamma_1 \sqrt{\frac{1}{m+n} \log \frac{|\mathcal{H}|}{\delta_K} \hat{\mathrm{V}}(h;\tilde{S}_K,M_k)}$.
```

7.3.2 Analysis

We have the following generalization error bound for Algorithm 11. Despite not querying for all labels, our algorithm achieves the same asymptotic bound as the one that queries labels for all online data.

Theorem 7.9. *Let* $M = \inf\{M' \geq 1 \mid \frac{2M'}{m+n} \log \frac{|\mathcal{H}|}{\delta_K} \geq \mathbb{P}(\frac{m+n}{mQ_0(X)+n} \geq M'/2)\}$ *be the final clipping threshold used in step 20. There is an absolute constant* $c_0 > 1$ *such that for any* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$\begin{aligned} l(\hat{h}) \leq & l(h^\star) + c_0 \sqrt{\mathbb{E} \frac{\mathbb{1}\{h^\star(X) \neq Y\}}{mQ_0(X)+n} \mathbb{1}\{\frac{m+n}{mQ_0(X)+n} \leq M\} \log \frac{|\mathcal{H}|}{\delta}} \\ & + c_0 \frac{M \log \frac{|\mathcal{H}|}{\delta}}{m+n} + c_0 \frac{M^2 \sqrt{\log \frac{|\mathcal{H}|}{\delta}}}{(m+n)^{\frac{3}{2}}}. \end{aligned}$$

Next, we analyze the number of labels queried by Algorithm 11 with the help of following definitions.

Definition 7.10. For any $t \geq 1, r > 0$, define the modified disagreement coefficient $\tilde{\theta}(r,t) := \frac{1}{r}\mathbb{P}\left(\mathrm{DIS}(B(h^\star, r)) \cap \left\{x : Q_0(x) \leq \frac{1}{t}\right\}\right)$. Define $\tilde{\theta} := \sup_{r>2\nu} \tilde{\theta}(r, \frac{2m}{n})$.

The modified disagreement coefficient $\tilde{\theta}(r,t)$ measures the probability of the intersection of two sets: the disagreement region for the r-ball around $h^\star$ and where the propensity score $Q_0(x)$ is smaller than $\frac{1}{t}$. It characterizes the size of the querying region of Algorithm 11. Note that the standard disagreement coefficient [Han07], which is widely used for analyzing DBAL in the classical active learning setting, can be written as $\theta(r) := \tilde{\theta}(r,1)$. Here, the modified disagreement coefficient modifies the standard definition to account for the reduction of the number of label queries due to the sample selection bias correction strategy: Algorithm 11 only queries examples on which $Q_0(x)$ is lower than some threshold, hence $\tilde{\theta}(r,t) \leq \theta(r)$. Moreover, our modified disagreement coefficient $\tilde{\theta}$ is always smaller than the modified disagreement coefficient of Chapter 6 (denoted by θ') which is used to analyze their algorithm.

Additionally, define $\alpha = \frac{m}{n}$ to be the size ratio of logged and online data, let $\tau_k = 2^k$, define $\xi = \min_{1 \le k \le K}\{M_k / \frac{m+n_k}{mq_0+n_k}\}$ to be the minimum ratio between the clipping threshold M_k and maximum MIS weight $\frac{m+n_k}{mq_0+n_k}$ ($\xi \le 1$ since $M_k \le \frac{m+n_k}{mq_0+n_k}$ by the choice of M_k), and define $\bar{M} = \max_{1 \le k \le K} M_k$ to be the maximum clipping threshold. Recall $q_0 = \inf_X Q_0(X)$.

The following theorem upper-bounds the number of label queries by Algorithm 11.

Theorem 7.11. *There is an absolute constant $c_1 > 1$ such that for any $\delta > 0$, with probability at least $1 - \delta$, the number of labels queried by Algorithm 11 is at most:*

$$c_1 \tilde{\theta} \cdot (n\nu + \sqrt{\frac{n\nu\xi}{\alpha q_0 + 1} \log \frac{|\mathcal{H}| \log n}{\delta}} + \frac{\bar{M}\xi \log n}{\sqrt{n\alpha}} \sqrt{\log \frac{|\mathcal{H}| \log n}{\delta}} + \frac{\xi \log n}{\alpha q_0 + 1} \log \frac{|\mathcal{H}| \log n}{\delta}).$$

7.3.3 Discussion

In this subsection, we compare the performance of the proposed algorithm and some alternatives to understand the effect of proposed techniques. The theoretical performance of learning algorithms is captured by label complexity, which is defined as the number of label queries required during the active learning phase to guarantee the test error of the output classifier to be at most $\nu + \varepsilon$ (here $\nu = l(h^\star)$ is the optimal error , and ε is the target excess error). This can be derived by combining the upper bounds on the error (Theorem 7.9) and the number of queries (Theorem 7.11).

- The label complexity is $\tilde{O}\left(\nu \tilde{\theta} \log |\mathcal{H}| \cdot \left(\frac{M}{\varepsilon(1+\alpha)} + \frac{1}{\varepsilon^2} \mathbb{E} \frac{\mathbb{1}\{h^\star(X) \neq Y\}}{1+\alpha Q_0(X)} \mathbb{1}\{\frac{1+\alpha}{1+\alpha Q_0(X)} \le M\}\right)\right)$ for

Algorithm 11. This is derived from Theorem 7.9, 7.11.

- The label complexity is $\tilde{O}\left(\nu\tilde{\theta}\log|\mathcal{H}|\cdot\left(\frac{1}{\varepsilon(1+\alpha q_0)}+\frac{1}{\varepsilon^2}\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{1+\alpha Q_0(X)}\right)\right)$ without clipping. This is derived by setting the final clipping threshold $M_K=\frac{1+\alpha}{1+\alpha q_0}$. It is worse since $\frac{1+\alpha}{1+\alpha q_0}\geq M$.

- The label complexity is $\tilde{O}\left(\nu\tilde{\theta}\log|\mathcal{H}|\cdot(\frac{1}{\varepsilon}+\frac{\nu}{\varepsilon^2})\frac{1}{1+\alpha q_0}\right)$ if regularizers are removed further. This is worse since $\frac{\nu}{1+\alpha q_0}\geq\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{1+\alpha Q_0(X)}$.

- The label complexity is $\tilde{O}\left(\nu\theta\log|\mathcal{H}|\cdot(\frac{1}{\varepsilon}+\frac{\nu}{\varepsilon^2})\frac{1}{1+\alpha q_0}\right)$ if we further remove the sample selection bias correction strategy. Here the standard disagreement coefficient θ is used ($\theta\geq\tilde{\theta}$).

- The label complexity is $\tilde{O}\left(\nu\theta\log|\mathcal{H}|\cdot\left(\frac{1}{\varepsilon(1+\alpha q_0)}+\frac{\nu(q_0+\alpha)}{\varepsilon^2(1+\alpha)^2q_0}\right)\right)$ if we further remove the MIS technique. It can be shown $\frac{q_0+\alpha}{(1+\alpha)^2q_0}\geq\frac{1}{1+\alpha q_0}$, so MIS gives a better label complexity bound.

- The label complexity is $\tilde{O}\left(\log|\mathcal{H}|\cdot\left(\frac{1}{\varepsilon(1+\alpha q_0)}+\frac{\nu(q_0+\alpha)}{\varepsilon^2(1+\alpha)^2q_0}\right)\right)$ if DBAL is further removed. Here, all n online examples are queried. This demonstrates that DBAL decreases the label complexity bound by a factor of $\nu\theta$ which is at most 1 by definition.

- Finally, the label complexity is $\tilde{O}\left(\nu\theta'\log|\mathcal{H}|\cdot\frac{\nu+\varepsilon}{\varepsilon^2}\frac{1}{1+\alpha q_0}\right)$ for Chapter 6, the only known algorithm in our setting. Here, $\theta'\geq\tilde{\theta}$, $\frac{\nu}{1+\alpha q_0}\geq\mathbb{E}\frac{\mathbb{1}\{h^\star(X)\neq Y\}}{1+\alpha Q_0(X)}$, and $\frac{1}{1+\alpha q_0}\geq\frac{M}{1+\alpha}$. Thus, the label complexity of the proposed algorithm is better than Chapter 6. This improvement is made possible by the second moment regularizer, the principled clipping technique, and thereby the improved sample selection bias correction strategy.

7.4 Acknowledgements

This chapter is based on the material submitted to Advances in Neural Information Processing Systems 2019 (Songbai Yan, Kamalika Chaudhuri and Tara Javidi, ”The Label Complexity of Active Learning from Observational Data”) [YCJ19]. The dissertation author is the co-primary investigator and co-author of this material.

Appendix A

Omitted Proofs for Chapter 4

A.1 Basic Lemmas

In this section, we present a few useful lemmas that serve as the basis of the proofs.

A.1.1 Basic Facts

We first collect a few useful facts for algebraic manipulations.

Lemma A.1. *If* $0 \leq x \leq 1 - \frac{1}{e}$*, then for any* $d \geq 1$*,* $(1 - \frac{x}{d})^{\frac{d}{2}} \geq e^{-x} \geq \frac{1}{2}$.

Lemma A.2. *Given* $a \in (0, \pi)$*, if* $x \in [0, a]$*, then* $\frac{\sin a}{a} x \leq \sin x \leq x$*.*

Lemma A.3. *If* $x \in [0, \pi]$*, then* $1 - \frac{x^2}{2} \leq \cos x \leq 1 - \frac{x^2}{5}$.

Lemma A.4. *Let* $\mathrm{B}(x, y) = \int_0^1 (1-t)^{x-1} t^{y-1} \, \mathrm{d}t$ *be the Beta function. Then* $\frac{2}{\sqrt{d-1}} \leq \mathrm{B}(\frac{1}{2}, \frac{d}{2}) \leq \frac{\pi}{\sqrt{d}}$.

A.1.2 Probability Inequalities

Lemma A.5 (Azuma's Inequality)**.** *Let $\{Y_t\}_{t=1}^m$ be a bounded submartingale difference sequence, that is, $\mathbb{E}[Y_t|Y_1,\ldots,Y_{t-1}] \geq 0$, and $|Y_t| \leq \sigma$. Then, with probability at least $1-\delta$,*

$$\sum_{t=1}^{m} Y_t \geq -\sigma\sqrt{2m\ln\frac{1}{\delta}}$$

Lemma A.6 (Concentration of Geometric Random Variables)**.** *Suppose $Z_1,\ldots,Z_n$ are iid geometric random variables with parameter p. Then,*

$$\mathbb{P}[Z_1+\ldots+Z_n > \frac{2n}{p}] \leq \exp(-\frac{n}{4})$$

Proof. Since $Z_1+\ldots+Z_n > \frac{2n}{p}$ implies that $Z_1+\ldots+Z_n \geq \lceil\frac{2n}{p}\rceil$ (as $Z_1+\ldots+Z_n$ is an integer), the left hand side is at most $\mathbb{P}[Z_1+\ldots+Z_n \geq \lceil\frac{2n}{p}\rceil]$.

Let $X_1,\ldots,X_{\lceil\frac{2n}{p}\rceil}$ be a sequence of iid Bernoulli(p) random variables. By standard relationship between Bernoulli random variables and geometric random variables, we have that

$$\mathbb{P}[Z_1+\ldots+Z_n \geq \lceil\frac{2n}{p}\rceil] = \mathbb{P}[X_1+\ldots+X_{\lceil\frac{2n}{p}\rceil-1} \leq n-1]$$

Note that $\mathbb{P}[X_1+\ldots+X_{\lceil\frac{2n}{p}\rceil-1} \leq n-1] \leq \mathbb{P}[X_1+\ldots+X_{\lceil\frac{2n}{p}\rceil} \leq n]$ since $X_{\lceil\frac{2n}{p}\rceil} \leq 1$. Applying Chernoff bound, the above probability is at most $\exp(-\lceil\frac{2n}{p}\rceil\cdot p\cdot\frac{1}{8}) \leq \exp(-\frac{n}{4})$. □

A.1.3 Properties of the Uniform Distribution over the Unit Sphere

Lemma A.7 (Marginal Density and Conditional Density). *If $(x_1, x_2, \ldots, x_d)$ is drawn from the uniform distribution over the unit sphere, then:*

1. *(x_1, x_2) has a density function of $p(z_1, z_2)$, where $p(z_1, z_2) = \frac{(1-z_1^2-z_2^2)^{\frac{d-4}{2}}}{\frac{2\pi}{d-2}}$.*
2. *Conditioned on $x_2 = b$, x_1 has a density function of $p_b(z)$, where $p_b(z) = \frac{(1-b^2-z^2)^{\frac{d-4}{2}}}{(1-b^2)^{\frac{d-3}{2}} \mathrm{B}(\frac{d-2}{2}, \frac{1}{2})}$.*
3. *x_1 has a density function of $p(z)$, where $p(z) = \frac{(1-z^2)^{\frac{d-3}{2}}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})}$.*

Lemma A.8. *Suppose x is drawn uniformly from the unit sphere, and $b \le \frac{1}{10\sqrt{d}}$. Then, $\mathbb{P}(x_1 \in [\frac{b}{2}, b]) \ge \frac{\sqrt{d}}{8\pi} b$.*

Proof.

$$\begin{aligned}
& \mathbb{P}(x_1 \in [\frac{b}{2}, b]) \\
= & \frac{\int_{b/2}^{b} (1-t^2)^{\frac{d-3}{2}} \,\mathrm{d}t}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})} \\
\ge & \frac{\frac{b}{2}(1-b^2)^{\frac{d-3}{2}}}{\frac{\pi}{\sqrt{d-1}}} \ge \frac{\sqrt{d}}{8\pi} b
\end{aligned}$$

where the first equality is from item 3 of Lemma A.7, giving the exact probability density function of x_1, the first inequality is from that $(1-t^2)^{\frac{d-3}{2}} \ge (1-b^2)^{\frac{d-3}{2}}$ when $t \in [b/2, b]$, and Lemma A.4 giving upper bound on $\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})$, and the second inequality is from Lemma A.1 and that $d - 1 \ge \frac{d}{2}$. □

Lemma A.9. *Suppose x is drawn uniformly from unit sphere restricted to the region $\{x : v \cdot x = \xi\}$, and u, v are unit vectors such that $\theta(u, v) = \theta \in [0, \frac{9}{10}\pi]$ and $0 \le \xi \le \frac{\theta}{4\sqrt{d}}$. Then,* *1.* $\mathbb{E}[u \cdot x] \leq \xi$.

2. $\mathbb{E}[(u \cdot x)^2] \leq \frac{5\theta^2}{d}$.

3. $\mathbb{E}[(u \cdot x)\mathbb{1}\{u \cdot x < 0\}] \leq \xi - \frac{\theta}{36\sqrt{d}}$.

Proof. By spherical symmetry, let $v = (0,1,0,\ldots,0)$ and $u = (\sin\theta, \cos\theta, 0, \ldots, 0)$ without loss of generality. Let $x = (x_1, \ldots, x_d)$.

1.

$$\begin{aligned}
& \mathbb{E}[u \cdot x] \\
= \; & \mathbb{E}[x_1 \sin\theta + x_2 \cos\theta | x_2 = \xi] \\
= \; & \mathbb{E}[x_1 | x_2 = \xi] \sin\theta + \xi \cos\theta \\
\leq \; & \xi
\end{aligned}$$

where the first two equalities are by algebra, the inequality follows from $\cos\theta \leq 1$ and $\mathbb{E}[x_1|x_2 = \xi] = 0$ since the conditional distribution of x_1 given $x_2 = \xi$ is symmetric around the origin.

2.

$$
\begin{aligned}
& \mathbb{E}[(u \cdot x)^2] \\
= \ & \mathbb{E}[(x_1 \sin\theta + x_2 \cos\theta)^2 | x_2 = \xi] \\
\leq \ & \mathbb{E}[2x_1^2 \sin^2\theta + 2x_2^2 \cos^2\theta | x_2 = \xi] \\
\leq \ & 2\mathbb{E}[x_1^2 | x_2 = \xi] \sin^2\theta + 2\xi^2 \\
\leq \ & 2\theta^2 \frac{\int_{-1}^{1} z^2 (1-z^2)^{\frac{d-4}{2}} \, \mathrm{d}z}{\mathrm{B}(\frac{d-2}{2}, \frac{1}{2})} + 2\xi^2 \\
= \ & 2\theta^2 \frac{\mathrm{B}(\frac{d-2}{2}, \frac{3}{2})}{\mathrm{B}(\frac{d-2}{2}, \frac{1}{2})} + 2\xi^2 \\
\leq \ & \frac{5\theta^2}{d}
\end{aligned}
$$

where the first equality is by definition of u, the first inequality is from algebra that $(A+B)^2 \leq 2A^2 + 2B^2$, the second inequality is from that $|\cos\theta| \leq 1$, the third inequality is from item 2 of Lemma A.7 and that $\sin\theta \leq \theta$, and the last inequality is from the fact that $\frac{\mathrm{B}(\frac{d-2}{2}, \frac{3}{2})}{\mathrm{B}(\frac{d-2}{2}, \frac{1}{2})} = \frac{1}{d-1} \leq \frac{2}{d}$, and $\xi^2 \leq \frac{\theta^2}{16d}$.

3.

$$
\begin{aligned}
& \mathbb{E}[(u \cdot x)\mathbb{1}\{u \cdot x < 0\}] \\
= \; & \mathbb{E}[(x_1 \sin\theta + x_2 \cos\theta)\mathbb{1}\left\{x_1 < -\xi \cot\theta\right\} | x_2 = \xi] \\
\leq \; & \mathbb{E}[x_1 \mathbb{1}\left\{x_1 < -\xi \cot\theta\right\} | x_2 = \xi] \sin\theta + \xi \\
= \; & \xi + \sin\theta \int_{-\sqrt{1-\xi^2}}^{-\xi\cot\theta} \frac{(1-\xi^2-x_1^2)^{\frac{d-4}{2}} x_1}{(1-\xi^2)^{\frac{d-3}{2}} \mathrm{B}(\frac{d-2}{2}, \frac{1}{2})} \mathrm{d}x_1 \\
= \; & \xi - \sin\theta \frac{\frac{2}{d-2}\left(1-\left(\frac{\xi}{\sin\theta}\right)^2\right)^{\frac{d-2}{2}}}{(1-\xi^2)^{\frac{d-3}{2}} \mathrm{B}(\frac{d-2}{2}, \frac{1}{2})} \\
\leq \; & \xi - \sin\theta \frac{2}{\pi\sqrt{d-2}} \left(1-\left(\frac{\xi}{\sin\theta}\right)^2\right)^{\frac{d-2}{2}} \\
\leq \; & \xi - \frac{\sin\theta}{\pi\sqrt{d}} \\
\leq \; & \xi - \frac{\theta}{36\sqrt{d}}
\end{aligned}
$$

where the first inequality is by algebra and $|\cos\theta| \leq 1$, the second equality is by item 2 of Lemma A.7, the third equality is by integration, the second inequality is from $(1-\xi^2)^{\frac{d-3}{2}} \leq 1$ and Lemma A.4 that $\mathrm{B}(\frac{d-2}{2}, \frac{1}{2}) \leq \frac{\pi}{\sqrt{d-2}}$, the third inequality follows by Lemma A.1 that $\left(1-\left(\frac{\xi}{\sin\theta}\right)^2\right)^{\frac{d-2}{2}} \geq \frac{1}{2}$, since $\xi \leq \frac{\theta}{4\sqrt{d}}$, and the last inequality follows from Lemma A.2 that $\sin\theta \geq \frac{5\theta}{18\pi}$ when $\theta \in [0, \frac{9}{10}\pi]$ and algebra.

□

Algorithm 12 Master Algorithm in the Bounded Noise Setting

```
Input: Labeler $O$, confidence $\delta$, noise upper bound $\eta$, sample schedule $\{m_k\}$, band width $\{b_k\}$.
Output: a halfspace $\hat{v}$ such that $\theta(\hat{v}, u) \le \frac{\pi}{4}$.
$v_0 \leftarrow (1, 0, \ldots, 0)$.
$v_+ \leftarrow$ ACTIVE-PERCEPTRON$(O, v_0, \frac{(1-2\eta)}{16}, \frac{\delta}{3}, \{m_k\}, \{b_k\})$.
$v_- \leftarrow$ ACTIVE-PERCEPTRON$(O, -v_0, \frac{(1-2\eta)}{16}, \frac{\delta}{3}, \{m_k\}, \{b_k\})$.
Define region $R := \{x : \mathrm{sign}(v_+ \cdot x) \neq \mathrm{sign}(v_- \cdot x)\}$.
$S \leftarrow$ Draw $\frac{8}{(1-2\eta)^2} \ln \frac{6}{\delta}$ iid examples from $D|_R$ and query their labels.
if $l_S(h_{v_+}) \le l_S(h_{v_-})$ then
    Return $v_+$
else
    Return $v_-$
end if
```

A.2 Acute Initialization

We show in this section that the angle between the initial vector v_0 and the underlying halfspace u can be assumed to be acute under the bounded noise model without loss of generality. To this end, we present Algorithm 12 that returns a halfspace that has angle at most $\frac{\pi}{4}$ with u, with constant overhead in label and time complexities. The techniques here are due to Appendix B of [ABL14]. This fact, in conjunction with Theorem 4.10, yields an active learning algorithm that learns the target halfspace unconditionally with a constant overhead of label and time complexities.

For the bounded noise setting, we construct Algorithm 12 as an initialization procedure. It runs Algorithm 2 ACTIVE-PERCEPTRON twice, taking a vector v_0 and its opposite direction $-v_0$ as initializers. Then it performs hypothesis testing using $\tilde{O}(\frac{1}{(1-2\eta)^2})$ labeled examples to identify a halfspace that has angle at most $\frac{\pi}{4}$ with u.

Theorem A.10. *Suppose Algorithm 12 has inputs labeler O that satisfies η-bounded noise condition with respect to u, confidence δ, sample schedule $m_k = \Theta\left(\frac{d}{(1-2\eta)^2}(\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta})\right)$, band width $\{b_k\}$ where $b_k = \tilde{\Theta}\left(\frac{2^{-k}(1-2\eta)}{\sqrt{d}}\right)$. Then, with probability at least $1-\delta$, the output $\hat{v}$ is* *such that $\theta(\hat{v},u) \le \frac{\pi}{4}$. Furthermore, (1) the total number of label queries to labeler O is at most $\tilde{O}\left(\frac{d}{(1-2\eta)^2}\right)$; (2) the total number of unlabeled examples drawn is $\tilde{O}\left(\frac{d}{(1-2\eta)^3}\right)$; (3) the algorithm runs in time $\tilde{O}\left(\frac{d^2}{(1-2\eta)^3}\right)$.*

Proof. Note that one of $\theta(v_0,u)$, $\theta(-v_0,u)$ is at most $\frac{\pi}{2}$. From Theorem 4.10 and union bound, we know that with probability at least $1-\frac{2\delta}{3}$, either $\theta(v_+,u) \le \frac{(1-2\eta)\pi}{16}$, or $\theta(v_-,u) \le \frac{(1-2\eta)\pi}{16}$.

Suppose without loss of generality, $\theta(v_+,u) \le \frac{(1-2\eta)\pi}{16}$. We consider two cases.

Case 1: $\theta(v_+,v_-) \le \pi/8$**.** By triangle inequality, $\theta(v_-,u) \le \theta(v_+,u) + \theta(v_+,v_-) \le \pi/4$. In this case, $\theta(v_+,u) \le \frac{\pi}{4}$ and $\theta(v_-,u) \le \frac{\pi}{4}$ holds simultaneously. Therefore, the returned vector $\hat{v}$ satisfies $\theta(\hat{v},u) \le \frac{\pi}{4}$.

Case 2: $\theta(v_+,v_-) > \pi/8$**.** In this case, $\mathbb{P}[x \in R] \ge 1/8$, thus,

$$\mathbb{P}_R[\text{sign}(v_+ \cdot x) \ne \text{sign}(u \cdot x)] \le \frac{\mathbb{P}[\text{sign}(v_+ \cdot x) \ne \text{sign}(u \cdot x)]}{\mathbb{P}[x \in R]} \le \frac{1-2\eta}{8} = \frac{1}{4}(\frac{1}{2}-\eta).$$

Meanwhile, $\mathbb{P}_R[\text{sign}(v_+ \cdot x) \ne y] \le \eta \mathbb{P}_R[\text{sign}(v_+ \cdot x) = \text{sign}(u \cdot x)] + \mathbb{P}_R[\text{sign}(v_+ \cdot x) \ne \text{sign}(u \cdot x)]$. Therefore,

$$\begin{aligned}
& \frac{1}{2} - \mathbb{P}_R[\text{sign}(v_+ \cdot x) \ne y] \\
\ge\ & (\frac{1}{2}-\eta)\mathbb{P}_R[\text{sign}(v_+ \cdot x) = \text{sign}(u \cdot x)] - \frac{1}{2}\mathbb{P}_R[\text{sign}(v_+ \cdot x) \ne \text{sign}(u \cdot x)] \\
\ge\ & (\frac{1}{2}-\eta) \cdot \frac{1}{2} - (\frac{1}{2}-\eta) \cdot \frac{1}{4} \\
\ge\ & \frac{1}{4}(\frac{1}{2}-\eta)
\end{aligned}$$

Since v_+ disagrees with v_- everywhere on R, $\mathbb{P}_R[\text{sign}(v_+ \cdot x) \neq y] + \mathbb{P}_R[\text{sign}(v_- \cdot x) \neq y] = 1$. Thus, $l_{D|_R}(h_{v_+}) \leq \frac{1}{2} - (\frac{1}{2} - \eta)\frac{1}{4}$ and $l_{D|_R}(h_{v_-}) \geq \frac{1}{2} + (\frac{1}{2} - \eta)\frac{1}{4}$. Therefore, by Hoeffding's Inequality, with probability at least $1 - \delta/3$,

$$l_S(v_+) < \frac{1}{2} < l_S(v_-)$$

therefore v_+ will be selected for $\hat{v}$. This shows that $\theta(\hat{v}, u) \leq \pi/4$.

In conclusion, by union bound, we have shown that with probability $1 - \delta$, $\theta(\hat{v}, u) \leq \frac{\pi}{4}$. The label complexity, unlabeled sample complexity, and time complexity of the algorithm follows immediately from Theorem 4.10. □

Appendix B

Omitted Proofs for Chapter 5

B.1 Technical lemmas

B.1.1 Concentration bounds

In this subsection, we define $Y_1, Y_2, \dots$ to be a sequence of i.i.d. random variables. Assume $Y_1 \in [-2,2]$, $\mathbb{E}Y_1 = 0$, $\mathrm{Var}(Y_1) = \sigma^2 \le 4$. Define $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right)$. It is easy to check $\mathbb{E}V_n = n\sigma^2$.

We need following two results from [RB16]

Lemma B.1. *([RB16], Theorem 2) Take any* $0 < \delta < 1$. *Then there is an absolute constant* D_0 *such that with probability at least* $1-\delta$*, for all n simultaneously,*

$$\left|\sum_{i=1}^n Y_i\right| \le D_0 \left(1 + \ln\frac{1}{\delta} + \sqrt{n\sigma^2 \left[\ln\ln\right]_+ (n\sigma^2) + n\sigma^2 \ln\frac{1}{\delta}}\right)$$

Lemma B.2. *([RB16], Lemma 3) Take any* $0 < \delta < 1$. *Then there is an absolute constant* K_0

such that with probability at least $1-\delta$*, for all n simultaneously,*

$$n\sigma^2 \leq K_0 \left(1+\ln\frac{1}{\delta}+\sum_{i=1}^{n} Y_i^2\right)$$

We note that Proposition 5.6 is immediate from Lemma B.1 since $\mathrm{Var}(Y_i) \leq 4$.

Lemma B.3. *Take any* $0<\delta<1$*. Then there is an absolute constant* K_3 *such that with probability at least* $1-\delta$*, for all* $n \geq \ln\frac{1}{\delta}$ *simultaneously,*

$$n\sigma^2 \leq K_3 \left(1+\ln\frac{1}{\delta}+V_n\right)$$

Proof. By Lemma B.2, with probability at least $1-\delta/2$, for all n,

$$n\sigma^2 \leq K_0 \left(\sum_{i=1}^{n} Y_i^2+\ln\frac{2}{\delta}+1\right) = K_0 \left(\frac{n-1}{n}V_n+\frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2+\ln\frac{2}{\delta}+1\right)$$

By Lemma B.1, with probability at least $1-\delta/2$, for all n,

$$\begin{aligned}
\frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 &< \frac{1}{n}\left(D_0\left(1+\ln\frac{2}{\delta}+\sqrt{n\sigma^2\left[\ln\ln\right]_+(n\sigma^2)+n\sigma^2\ln\frac{2}{\delta}}\right)\right)^2 \\
&= \frac{D_0^2}{n}\left(1+\ln\frac{2}{\delta}\right)^2+D_0^2\sigma^2\left[\ln\ln\right]_+(n\sigma^2)+D_0^2\sigma^2\ln\frac{2}{\delta} \\
&\quad +2D_0^2\left(1+\ln\frac{2}{\delta}\right)\sqrt{\frac{\sigma^2\left[\ln\ln\right]_+(n\sigma^2)+\sigma^2\ln\frac{2}{\delta}}{n}} \\
&\leq K_1\left(1+\ln\frac{1}{\delta}+\left[\ln\ln\right]_+(n\sigma^2)\right)
\end{aligned}$$

for some absolute constant K_1. The last inequality follows by $n \geq \ln\frac{1}{\delta}$.

Thus, by a union bound, with probability at least $1-\delta$, for all n, $n\sigma^2 \leq K_0 V_n + K_0(K_1 + 2)\ln\frac{1}{\delta} + K_0 K_1 \left[\ln\ln\right]_+ (n\sigma^2) + K_0(K_1+3)$.

Let $K_2 > 0$ be an absolute constant such that $\forall x \geq K_2$, $K_0 K_1 \left[\ln\ln\right]_+ x \leq \frac{x}{2}$.

Now if $n\sigma^2 \geq K_2$, then $n\sigma^2 \leq K_0 V_n + K_0(K_1+2)\ln\frac{1}{\delta} + \frac{n\sigma^2}{2} + K_0(K_1+3)$, and thus

$$n\sigma^2 \leq 2K_0 V_n + 2K_0(K_1+2)\ln\frac{1}{\delta} + 2K_0(K_1+3) + K_2 \tag{B.1}$$

If $n\sigma^2 \leq K_2$, clearly (B.1) holds. This concludes the proof. □

We note that Proposition 5.7 is immediate by applying above lemma to Lemma B.1.

Lemma B.4. *Take any $\delta, n > 0$. Then with probability at least $1-\delta$,*

$$V_n \leq 4n\sigma^2 + 8\ln\frac{1}{\delta}$$

Proof. Applying Bernstein's Inequality to Y_i^2, and noting that $\mathrm{Var}(Y_i^2) \leq 4\sigma^2$ since $|Y_i| \leq 2$, we have with probability at least $1-\delta$,

$$\begin{aligned}\sum_{i=1}^n Y_i^2 &\leq \frac{4}{3}\ln\frac{1}{\delta} + n\sigma^2 + \sqrt{8n\sigma^2\ln\frac{1}{\delta}} \\ &\leq 4\ln\frac{1}{\delta} + 2n\sigma^2\end{aligned}$$

The last inequality follows by the fact that $\sqrt{4ab} \leq a+b$.

The desired result follows by noting that $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right) \leq 2\sum_{i=1}^n Y_i^2$. □

B.1.2 Bounds of distances among probability distributions

Lemma B.5. *If P, Q are two probability distributions on a countable support $\mathcal{X}$, then*

$$d_{KL}\left(P \parallel Q\right) \leq \sum_x \frac{\left(P(x) - Q(x)\right)^2}{Q(x)}$$

Proof.

$$\begin{aligned} d_{\mathrm{KL}}\left(P \parallel Q\right) &= \sum_x P(x) \ln \frac{P(x)}{Q(x)} \\ &\leq \sum_x P(x) \left(\frac{P(x)}{Q(x)} - 1\right) \\ &= \sum_x \frac{\left(P(x) - Q(x)\right)^2}{Q(x)} \end{aligned}$$

The first inequality follows by $\ln x \leq x - 1$. The second equality follows by

$$\sum_x P(x) \left(\frac{P(x)}{Q(x)} - 1\right) = \sum_x \left(\frac{P^2(x) - P(x)Q(x)}{Q(x)} - P(x) + Q(x)\right) = \sum_x \frac{\left(P(x) - Q(x)\right)^2}{Q(x)}.$$

□

Define

$$
\begin{aligned}
P_0\left(Y=\perp|x\right) &= 1-\left|x-\frac{1}{2}\right|^{\alpha}\\
P_0\left(Y=0|x\right) &= \begin{cases}\left(x-\frac{1}{2}\right)^{\alpha}\left(1-\left(x-\frac{1}{2}\right)^{\beta}\right)/2 & x>\frac{1}{2}\\ \left(\frac{1}{2}-x\right)^{\alpha}\left(1+\left(\frac{1}{2}-x\right)^{\beta}\right)/2 & x\leq\frac{1}{2}\end{cases}\\
P_0\left(Y=1|x\right) &= \begin{cases}\left(x-\frac{1}{2}\right)^{\alpha}\left(1+\left(x-\frac{1}{2}\right)^{\beta}\right)/2 & x>\frac{1}{2}\\ \left(\frac{1}{2}-x\right)^{\alpha}\left(1-\left(\frac{1}{2}-x\right)^{\beta}\right)/2 & x\leq\frac{1}{2}\end{cases}
\end{aligned}
$$

and

$$
\begin{aligned}
P_1\left(Y=\perp|x\right) &= 1-\left|x-\varepsilon-\frac{1}{2}\right|^{\alpha}\\
P_1\left(Y=0|x\right) &= \begin{cases}\left(x-\varepsilon-\frac{1}{2}\right)^{\alpha}\left(1-\left(x-\varepsilon-\frac{1}{2}\right)^{\beta}\right)/2 & x>\varepsilon+\frac{1}{2}\\ \left(\varepsilon+\frac{1}{2}-x\right)^{\alpha}\left(1+\left(\varepsilon+\frac{1}{2}-x\right)^{\beta}\right)/2 & x\leq\varepsilon+\frac{1}{2}\end{cases}\\
P_1\left(Y=1|x\right) &= \begin{cases}\left(x-\varepsilon-\frac{1}{2}\right)^{\alpha}\left(1+\left(x-\varepsilon-\frac{1}{2}\right)^{\beta}\right)/2 & x>\varepsilon+\frac{1}{2}\\ \left(\varepsilon+\frac{1}{2}-x\right)^{\alpha}\left(1-\left(\varepsilon+\frac{1}{2}-x\right)^{\beta}\right)/2 & x\leq\varepsilon+\frac{1}{2}\end{cases}
\end{aligned}
$$

Lemma B.6. *Let P_0, P_1 be the distributions defined above. If $x\in[0,1]$, $\varepsilon\leq\min\{(\frac{1}{2})^{1/\beta},(\frac{4}{5})^{1/\alpha},\frac{1}{4}\}$, then*

$$\sum_y\frac{\left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{P_0(Y=y|x)+P_1(Y=y|x)}=O\left(\varepsilon^{\alpha}+\varepsilon^2\right)\tag{B.2}$$

Proof. By symmetry, it suffices to show for $0\leq x\leq\frac{1+\varepsilon}{2}$. Let $t=\frac{1}{2}+\varepsilon-x$.

We first show (B.2) holds for $\frac{\varepsilon}{2} \le t \le \varepsilon$ (i.e. $\frac{1}{2} \le x \le \frac{1+\varepsilon}{2}$).

We claim $\min_y \left(P_0(Y=y|X=t)+P_1(Y=y|X=t)\right) \ge \frac{1}{2}\left(\frac{\varepsilon}{2}\right)^{\alpha}$. This is because:

- $P_0(Y=\perp|X=t)+P_1(Y=\perp|X=t) = 1-(\varepsilon-t)^{\alpha}+1-t^{\alpha} \ge 2-2\varepsilon^{\alpha} \ge \frac{1}{2}\left(\frac{\varepsilon}{2}\right)^{\alpha}$ where the last inequality follows by $\varepsilon \le \left(\frac{4}{5}\right)^{1/\alpha}$;
- $2(P_0(Y=0|X=t)+P_1(Y=0|X=t)) = (\varepsilon-t)^{\alpha}(1-(\varepsilon-t)^{\beta})+t^{\alpha}(1+t^{\beta}) \ge t^{\alpha}(1+t^{\beta}) \ge (\frac{\varepsilon}{2})^{\alpha}$. Therefore, $P_0(Y=0|X=t)+P_1(Y=0|X=t) \ge \frac{1}{2}(\frac{\varepsilon}{2})^{\alpha}$.
- Similarly, $P_0(Y=1|X=t)+P_1(Y=1|X=t) \ge \frac{1}{2}\left(\frac{\varepsilon}{2}\right)^{\alpha}$.

Besides,

$$
\begin{aligned}
&\sum_y \left(P_0(Y=y|X=t)-P_1(Y=y|X=t)\right)^2 \\
=&\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\frac{1}{4}\left(t^{\alpha}\left(1-t^{\beta}\right)-(\varepsilon-t)^{\alpha}\left(1+(\varepsilon-t)^{\beta}\right)\right)^2 \\
&+\frac{1}{4}\left(t^{\alpha}\left(1+t^{\beta}\right)-(\varepsilon-t)^{\alpha}\left(1-(\varepsilon-t)^{\beta}\right)\right)^2 \\
=&\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\frac{1}{4}\left(t^{\alpha}-(\varepsilon-t)^{\alpha}-t^{\alpha+\beta}-(\varepsilon-t)^{\alpha+\beta}\right)^2 \\
&+\frac{1}{4}\left(t^{\alpha}-(\varepsilon-t)^{\alpha}+t^{\alpha+\beta}+(\varepsilon-t)^{\alpha+\beta}\right)^2 \\
\overset{(a)}{\le}&\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha+\beta}+(\varepsilon-t)^{\alpha+\beta}\right)^2 \\
&+\frac{1}{2}\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha+\beta}+(\varepsilon-t)^{\alpha+\beta}\right)^2 \\
=&2\left(t^{\alpha}-(\varepsilon-t)^{\alpha}\right)^2+\left(t^{\alpha+\beta}+(\varepsilon-t)^{\alpha+\beta}\right)^2 \\
\le&2\varepsilon^{2\alpha}+4\varepsilon^{2\alpha+2\beta} \\
\le&6\varepsilon^{2\alpha}
\end{aligned}
$$

where (a) follows by the inequality $(a+b)^2 \le 2a^2+2b^2$ for any a, b.

Therefore, we get $\sum_y \frac{\left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{P_0(Y=y|x)+P_1(Y=y|x)} \leq \frac{\sum_y \left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{\min_y \left(P_0(Y=y|x)+P_1(Y=y|x)\right)} \leq 12 * 2^\alpha \varepsilon^\alpha$ when $\frac{1}{2} \leq x \leq \frac{1+\varepsilon}{2}$.

Next, We show (B.2) holds for $\varepsilon \leq t \leq \frac{1}{2}+\varepsilon$ (i.e. $0 \leq x \leq \frac{1}{2}$). We will show for $Y = \perp, 1, 0$, $\frac{\left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{P_0(Y=y|x)+P_1(Y=y|x)} = O\left(\varepsilon^\alpha + \varepsilon^2\right)$.

For $Y = \perp$, for the denominator,

$$P_0(Y = \perp |X = t) + P_1(Y = \perp |X = t) = 2 - t^\alpha - (t-\varepsilon)^\alpha \geq 2 - \left(\frac{3}{4}\right)^\alpha - \left(\frac{1}{2}\right)^\alpha$$

For the numerator,

$$\left(P_0(Y = \perp |X = t) - P_1(Y = \perp |X = t)\right)^2 = \left(t^\alpha - (t-\varepsilon)^\alpha\right)^2 = t^{2\alpha}\left(1 - \left(1 - \frac{\varepsilon}{t}\right)^\alpha\right)^2$$

By Lemma B.8, if $\alpha \geq 1$, $t^{2\alpha}\left(1 - \left(1 - \frac{\varepsilon}{t}\right)^\alpha\right)^2 \leq t^{2\alpha}\left(\alpha\frac{\varepsilon}{t}\right)^2 = t^{2\alpha-2}(\alpha\varepsilon)^2 = O\left(\varepsilon^2\right)$. If $0 \leq \alpha \leq 1$, $t^{2\alpha}\left(1 - \left(1 - \frac{\varepsilon}{t}\right)^\alpha\right)^2 \leq t^{2\alpha}\left(\frac{\varepsilon}{t}\right)^2 = t^{2\alpha-2}\varepsilon^2 \leq \varepsilon^{2\alpha}$.

Thus, we have $\frac{\left(P_0(Y=\perp|x)-P_1(Y=\perp|x)\right)^2}{P_0(Y=\perp|x)+P_1(Y=\perp|x)} = O\left(\varepsilon^{2\alpha} + \varepsilon^2\right)$.

For $Y = 1$, for the denominator,

$$\begin{aligned} 2\left(P_0(Y=1|X=t) + P_1(Y=1|X=t)\right) &= t^\alpha\left(1 - t^\beta\right) + (t-\varepsilon)^\alpha\left(1 - (t-\varepsilon)^\beta\right) \\ &\geq t^\alpha\left(1 - t^\beta\right) \\ &\geq t^\alpha\left(1 - \left(\frac{3}{4}\right)^\beta\right) \end{aligned}$$

For the numerator,

$$\begin{aligned}
&\left(P_0(Y=1|X=t)-P_1(Y=1|X=t)\right)^2\\
=&\frac{1}{4}\left(t^{\alpha}\left(1-t^{\beta}\right)-(t-\varepsilon)^{\alpha}\left(1-(t-\varepsilon)^{\beta}\right)\right)^2\\
\le&\frac{1}{2}\left(t^{\alpha}-(t-\varepsilon)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha+\beta}-(t-\varepsilon)^{\alpha+\beta}\right)^2\\
=&\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha}\right)^2+\frac{1}{2}t^{2\alpha+2\beta}\left(1-(1-\frac{\varepsilon}{t})^{\alpha+\beta}\right)^2\\
\le&\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha}\right)^2+\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha+\beta}\right)^2
\end{aligned}$$

If $\alpha \ge 1$, by Lemma B.8, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha}\right)^2+\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\alpha\frac{\varepsilon}{t}\right)^2 + \frac{1}{2}t^{2\alpha}\left((\alpha+\beta)\frac{\varepsilon}{t}\right)^2 = \left(\frac{1}{2}\alpha^2+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\varepsilon^2$. Thus,

$$\frac{\left(P_0(Y=1|x)-P_1(Y=1|x)\right)^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le \left(\frac{1}{2}\alpha^2+\frac{1}{2}(\alpha+\beta)^2\right)t^{\alpha-2}\varepsilon^2/\left(1-\left(\frac{3}{4}\right)^{\beta}\right)$$

which is $O(\varepsilon^2)$ if $\alpha \ge 2$ and $O(\varepsilon^{\alpha})$ if $\alpha \le 2$.

If $\alpha \le 1$ and $\alpha+\beta \ge 1$, by Lemma B.8, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha}\right)^2+\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\frac{\varepsilon}{t}\right)^2+\frac{1}{2}t^{2\alpha}\left((\alpha+\beta)\frac{\varepsilon}{t}\right)^2 = \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\varepsilon^2 \le \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\varepsilon^2$. Thus, $\frac{\left(P_0(Y=1|x)-P_1(Y=1|x)\right)^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{\alpha-2}\varepsilon^2/\left(1-\left(\frac{3}{4}\right)^{\beta}\right) = O(\varepsilon^{\alpha})$.

If $\alpha \le 1$, $\alpha+\beta \le 1$, by Lemma B.8, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha}\right)^2+\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\varepsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\frac{\varepsilon}{t}\right)^2+\frac{1}{2}t^{2\alpha}\left(\frac{\varepsilon}{t}\right)^2 = t^{2\alpha-2}\varepsilon^2$. Thus, $\frac{\left(P_0(Y=1|x)-P_1(Y=1|x)\right)^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le t^{\alpha-2}\varepsilon^2/\left(1-\left(\frac{3}{4}\right)^{\beta}\right) = O(\varepsilon^{\alpha})$.

Therefore, we have $\frac{\left(P_0(Y=1|x)-P_1(Y=1|x)\right)^2}{P_0(Y=1|x)+P_1(Y=1|x)} = O\left(\varepsilon^{\alpha}+\varepsilon^2\right)$.

Likewise, we can get $\frac{\left(P_0(Y=0|x)-P_1(Y=0|x)\right)^2}{P_0(Y=0|x)+P_1(Y=0|x)} = O\left(\varepsilon^{\alpha}+\varepsilon^2\right)$.

Thus, we prove $\sum_y \frac{\left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{P_0(Y=y|x)+P_1(Y=y|x)} = O\left(\varepsilon^\alpha + \varepsilon^2\right)$ when $x \leq \frac{1}{2}$. This concludes the proof. □

B.1.3 Other lemmas

Lemma B.7. *([RR11b], Lemma 4) For sufficiently large $d > 0$, there is a subset $M \subset \{0,1\}^d$ with following properties: (i) $|M| \geq 2^{d/48}$; (ii) $\left\|v - v'\right\|_0 > \frac{d}{12}$ for any two distinct $v, v' \in M$; (iii) for any $i = 1, \ldots, d$, $\frac{1}{24} \leq \frac{1}{M} \sum_{v \in M} v_i \leq \frac{3}{24}$.*

Lemma B.8. *If $x \leq 1, r \geq 1$, then $(1-x)^r \geq 1 - rx$ and $1-(1-x)^r \leq rx$.*

If $0 \leq x \leq 1, 0 \leq r \leq 1$, then $(1-x)^r \geq \frac{1-x}{1-x+rx}$ and $1-(1-x)^r \leq \frac{rx}{1-(1-r)x} \leq x$.

Inequalities above are know as Bernoulli's inequalities. One proof can be found in [LY13].

Lemma B.9. *Suppose ε, τ are positive numbers and $\delta \leq \frac{1}{2}$. Suppose $\{Z_i\}_{i=1}^\infty$ is a sequence of i.i.d random variables bounded by 1, $\mathbb{E}Z_i \geq \tau\varepsilon$, and $Var(Z_i) = \sigma^2 \leq 2\varepsilon$. Define $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Z_i - \frac{1}{n}\left(\sum_{i=1}^n Z_i\right)^2\right)$, $q_n = q(n, V_n, \delta)$ as Procedure 8. If $n \geq \frac{\eta}{\tau\varepsilon} \ln \frac{1}{\delta}$ for some sufficiently large number η (to be specified in the proof), then with probability at least $1-\delta$, $\frac{q_n}{n} - \mathbb{E}Z_i \leq -\tau\varepsilon/2$.*

Proof. By Lemma B.4, with probability at least $1-\delta$, $V_n \leq 4n\sigma^2 + 8\ln\frac{1}{\delta}$, which implies

$$q_n \leq D_1 \left(1 + \ln\frac{1}{\delta} + \sqrt{\left(4n\sigma^2 + 9\ln\frac{1}{\delta} + 1\right)\left([\ln\ln]_+ (4n\sigma^2 + 9\ln\frac{1}{\delta} + 1) + \ln\frac{1}{\delta}\right)}\right)$$

We denote the RHS by q.

On this event, we have

$$
\begin{aligned}
\frac{q_n}{n} - \mathbb{E}Z_i &\le \frac{q}{n} - \tau\varepsilon \\
&= \tau\varepsilon\left(\frac{q}{n\tau\varepsilon} - 1\right) \\
&\overset{(a)}{\le} \tau\varepsilon\left(\frac{2D_1}{\eta} + \frac{D_1}{\eta\ln\frac{1}{\delta}}\sqrt{\frac{9\eta}{\tau}\ln\frac{1}{\delta}\left([\ln\ln]_+(\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \ln\frac{1}{\delta}\right)} - 1\right) \\
&= \tau\varepsilon\left(\frac{2D_1}{\eta} + D_1\sqrt{\frac{9}{\eta\tau\ln\frac{1}{\delta}}[\ln\ln]_+(\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \frac{9}{\eta\tau}} - 1\right)
\end{aligned}
$$

where (a) follows from $\frac{q}{n}$ being monotonically decreasing with respect to n. By choosing η sufficiently large, we have $\frac{2D_1}{\eta} + D_1\sqrt{\frac{9}{\eta\tau\ln\frac{1}{\delta}}[\ln\ln]_+(\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \frac{9}{\eta\tau}} - 1 \le -\frac{1}{2}$, and thus $\frac{q_n}{n} - \mathbb{E}Z_i \le -\tau\varepsilon/2$. □

Appendix C

Omitted Proofs for Chapter 6

C.1 Preliminaries

C.1.1 Summary of Key Notations

Data Partitions $T_k = \{(X_t, Y_t, Z_t)\}_{t=m+n_1+\cdots+n_{k-1}+1}^{t=m+n_1+\cdots+n_k}$ $(1 \le k \le K)$ is the online data collected in k-th iteration of size $n_k = 2^{k-1}$. $n = n_1 + \cdots + n_K$, $\alpha = 2m/3n$. We define $n_0 = 0$. $T_0 = \{(X_t, Y_t, Z_t)\}_{t=1}^{t=m}$ is the logged data and is partitioned into $K+1$ parts $T_0^{(0)}, \cdots, T_0^{(K)}$ of sizes $m_0 = m/3, m_1 = \alpha n_1, m_2 = \alpha n_2, \cdots, m_K = \alpha n_K$. $S_k = T_0^{(k)} \cup T_k$.

Recall that $\tilde{S}_k$ and $\tilde{T}_k$ contain inferred labels while S_k and T_k are sets of examples with original labels. The algorithm only observes $\tilde{S}_k$ and $\tilde{T}_k$.

For $(X, Z) \in T_k$ $(0 \le k \le K)$, $Q_k(X) = \mathbb{P}(Z = 1 \mid X)$.

Disagreement Regions The candidate set C_k and its disagreement region D_k are defined in Algorithm 10. $\hat{h}_k = \arg\min_{h \in C_k} l(h, \tilde{S}_k)$. $\nu = l(h^\star)$.

$B(h,r) := \{h' \in \mathcal{H} \mid \rho(h,h') \leq r\}$, $\mathrm{DIS}(C) := \{x \in \mathcal{X} \mid \exists h_1 \neq h_2 \in C \text{ s.t. } h_1(x) \neq h_2(x)\}$.

$S(A,\alpha) = \bigcup_{A' \subseteq A} \left(A' \cap \left\{ x : Q_0(x) \leq \inf_{x \in A'} Q_0(x) + \frac{1}{\alpha} \right\} \right)$.

$\tilde{\theta}(r_0, \alpha) = \sup_{r > r_0} \frac{1}{r} \mathbb{P}(S(\mathrm{DIS}(B(h^\star, r)), \alpha))$.

$\mathrm{DIS}_0 = \mathcal{X}$. For $k = 1, \ldots, K$, $\mathrm{DIS}_k = \mathrm{DIS}(B(h^\star, 2\nu + \varepsilon_k))$, and

$$\varepsilon_k = \gamma_2 \sup_{x \in \mathrm{DIS}_{k-1}} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_{k-1} Q_0(x) + n_{k-1}} + \gamma_2 \sqrt{\sup_{x \in \mathrm{DIS}_{k-1}} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_{k-1} Q_0(x) + n_{k-1}} l(h^\star)}.$$

Other Notations $\rho(h_1,h_2) = \mathbb{P}(h_1(X) \neq h_2(X))$, $\rho_S(h_1,h_2) = \frac{1}{|S|} \sum_{X \in S} \mathbb{1}\{h_1(X) \neq h_2(X)\}$. For $k \geq 0$, $\sigma(k,\delta) = \sup_{x \in D_k} \frac{\log(|\mathcal{H}|/\delta)}{m_k Q_0(x) + n_k}$, $\delta_k = \frac{\delta}{(k+1)(k+2)}$. $\xi_k = \inf_{x \in D_k} Q_0(x)$. $\zeta = \sup_{x \in \mathrm{DIS}_1} \frac{1}{\alpha Q_0(x) + 1}$.

C.1.2 Elementary Facts

Proposition C.1. *Suppose $a, c \geq 0, b \in \mathbb{R}$. If $a \leq b + \sqrt{ca}$, then $a \leq 2b + c$.*

Proof. Since $a \leq b + \sqrt{ca}$, $\sqrt{a} \leq \frac{\sqrt{c} + \sqrt{c+4b}}{2} \leq \sqrt{\frac{c+c+4b}{2}} = \sqrt{c+2b}$ where the second inequality follows from the Root-Mean Square-Arithmetic Mean inequality. Thus, $a \leq 2b + c$. □

C.1.3 Facts on Disagreement Regions and Candidate Sets

Lemma C.2. *For any $k = 0, \ldots, K$, any $x \in \mathcal{X}$, any $h_1, h_2 \in C_k$, $\frac{\mathbb{1}\{h_1(x) \neq h_2(x)\}}{m_k Q_0(X) + n_k Q_k(X)} \leq \sup_{x'} \frac{\mathbb{1}\{x' \in D_k\}}{m_k Q_0(x') + n_k}$.*

Proof. The $k = 0$ case is obvious since $D_0 = \mathcal{X}$ and $n_0 = 0$.

For $k > 0$, since $\mathrm{DIS}(C_k) = D_k$, $\mathbb{1}\{h_1(x) \neq h_2(x)\} \leq \mathbb{1}\{x \in D_k\}$, and consequently $\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_k Q_0(X)+n_k Q_k(X)} \leq \frac{\mathbb{1}\{x\in D_k\}}{m_k Q_0(X)+n_k Q_k(X)}$.

For any x, if $Q_0(x) \leq \xi_k + 1/\alpha$, then $Q_k(x) = 1$, so $\frac{\mathbb{1}\{x\in D_k\}}{m_k Q_0(X)+n_k Q_k(X)} = \frac{\mathbb{1}\{x\in D_k\}}{m_k Q_0(x)+n_k} \leq \sup_{x'} \frac{\mathbb{1}\{x'\in D_k\}}{m_k Q_0(x')+n_k}$.

If $Q_0(x) > \xi_k + 1/\alpha$, then $Q_k(x) = 0$, and consequently $\frac{\mathbb{1}\{x\in D_k\}}{m_k Q_0(X)+n_k Q_k(X)} = \frac{\mathbb{1}\{x\in D_k\}}{m_k Q_0(x)} \leq \frac{\mathbb{1}\{x\in D_k\}}{m_k \xi_k+n_k} \leq \sup_{x'} \frac{\mathbb{1}\{x'\in D_k\}}{m_k Q_0(x')+n_k}$ where the first inequality follows from the fact that $Q_0(x) > \xi_k + 1/\alpha$ implies $m_k Q_0(x) > m_k \xi_k + n_k$ □

Lemma C.3. *For any $k = 0, \dots, K$, if $h_1, h_2 \in C_k$, then $l(h_1, S_k) - l(h_2, S_k) = l(h_1, \tilde{S}_k) - l(h_2, \tilde{S}_k)$.*

Proof. For any $(X_t, Y_t, Z_t) \in S_t$ that $Z_t = 1$, if $X_t \in \mathrm{DIS}(C_k)$, then $Y_t = \tilde{Y}_t$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\}$. If $X_t \notin \mathrm{DIS}(C_k)$, then $h_1(X_t) = h_2(X_t)$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\} = 0$. □

The following lemma is immediate from definition.

Lemma C.4. *For any $r \geq 2\nu$, any $\alpha \geq 1$, $\mathbb{P}(S(DIS(B(h^\star, r)), \alpha)) \leq r\tilde{\theta}(r, \alpha)$.*

C.1.4 Facts on Multiple Importance Sampling Estimators

We recall that $\{(X_t, Y_t)\}_{t=1}^{n_0+n}$ is an i.i.d. sequence. Moreover, the following fact is immediate by our construction that $S_0, \cdots, S_K$ are disjoint and that Q_k is determined by $S_0, \cdots, S_{k-1}$.

Fact C.5. *For any $0 \leq k \leq K$, conditioned on Q_k, examples in S_k are independent, and examples in T_k are i.i.d.. Besides, for any $0 < k \leq K$, Q_k, $T_0^{(k)}, \dots, T_0^{(K)}$ are independent.*

Unless otherwise specified, all probabilities and expectations are over the random draw of all random variables (including $S_0, \cdots, S_K$, $Q_1, \cdots, Q_K$).

The following lemma shows multiple importance estimators are unbiased.

Lemma C.6. *For any $h \in \mathcal{H}$, any $0 \leq k \leq K$, $\mathbb{E}[l(h,S_k)] = l(h)$.*

The above lemma is immediate from the following lemma.

Lemma C.7. *For any $h \in \mathcal{H}$, any $0 \leq k \leq K$, $\mathbb{E}[l(h,S_k) \mid Q_k] = l(h)$.*

Proof. The $k = 0$ case is obvious since $S_0 = T_0^{(0)}$ is an i.i.d. sequence and $l(h,S_k)$ reduces to a standard importance sampling estimator. We only show proof for $k > 0$.

Recall that $S_k = T_0^{(k)} \cup T_k$, and that $T_0^{(k)}$ and T_k are two i.i.d. sequences conditioned Q_k. We denote the conditional distributions of $T_0^{(k)}$ and T_k given Q_k by P_0 and P_k respectively. We have

$$\begin{aligned}
&\mathbb{E}[l(h,S_k) \mid Q_k] \\
= \;& \mathbb{E}\left[\sum_{(X,Y,Z)\in T_0^{(k)}} \frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + \mathbb{E}\left[\sum_{(X,Y,Z)\in T_k} \frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] \\
= \;& m_k \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + n_k \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right]
\end{aligned}$$

where the second equality follows since $T_0^{(k)}$ and T_k are two i.i.d. sequences given Q_k with sizes m_k and n_k respectively.

Now,

$$\begin{aligned}
\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] &= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid X, Q_k\right] \mid Q_k\right] \\
&= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_0(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid X, Q_k\right] \mid Q_k\right] \\
&= \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_0(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right]
\end{aligned}$$

where the second equality uses the definition $\mathbb{P}_{P_0}(Z \mid X) = Q_0(X)$ and the fact that $T_0^{(k)}$ and Q_k are independent.

Similarly, we have $\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Z}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] = \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Q_k(X)}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right]$.

Therefore,

$$\begin{aligned}
& m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Z}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] + n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Z}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] \\
= \quad & m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Q_0(X)}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] + n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Q_k(X)}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] \\
= \quad & \mathbb{E}_{P_0}\left[\mathbb{1}\{h(X)\neq Y\}\frac{m_kQ_0(X)+n_kQ_k(X)}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right] \\
= \quad & \mathbb{E}_D\left[\mathbb{1}\{h(X)\neq Y\}\right] = l(h)
\end{aligned}$$

where the second equality uses the fact that distribution of (X,Y) according to P_0 is the same as that according to P_k, and the third equality follows by algebra and Fact C.5 that Q_k is independent with $T_0^{(k)}$. □

The following lemma will be used to upper-bound the variance of the multiple importance sampling estimator.

Lemma C.8. *For any $h_1, h_2 \in \mathcal{H}$, any $0 \leq k \leq K$,*

$$\mathbb{E}\left[\sum_{(X,Y,Z)\in S_k}\left(\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{m_kQ_0(X)+n_kQ_k(X)}\right)^2 \mid Q_k\right] \leq \rho(h_1,h_2)\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}.$$

Proof. We only show proof for $k > 0$. The $k = 0$ case can be proved similarly.

We denote the conditional distributions of $T_0^{(k)}$ and T_k given Q_k by P_0 and P_k respectively.

Now, similar to the proof of Lemma C.7, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{(X,Y,Z)\in S_k}\left(\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{m_kQ_0(X)+n_kQ_k(X)}\right)^2 \mid Q_k\right]\\
=&\sum_{(X,Y,Z)\in S_k}\mathbb{E}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{\left(m_kQ_0(X)+n_kQ_k(X)\right)^2}\mid Q_k\right]\\
=&m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{\left(m_kQ_0(X)+n_kQ_k(X)\right)^2}\mid Q_k\right]+n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{\left(m_kQ_0(X)+n_kQ_k(X)\right)^2}\mid Q_k\right]\\
=&m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Q_0(X)}{\left(m_kQ_0(X)+n_kQ_k(X)\right)^2}\mid Q_k\right]+n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Q_k(X)}{\left(m_kQ_0(X)+n_kQ_k(X)\right)^2}\mid Q_k\right]\\
=&\mathbb{E}_{P_0}\left[\mathbb{1}\{h_1(X)\neq h_2(X)\}\frac{m_kQ_0(X)+n_kQ_k(X)}{(m_kQ_0(X)+n_kQ_k(X))^2}\mid Q_k\right]\\
=&\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}}{m_kQ_0(X)+n_kQ_k(X)}\mid Q_k\right]\\
\leq&\mathbb{E}_{P_0}\left[\mathbb{1}\{h_1(X)\neq h_2(X)\}\mid Q_k\right]\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}\\
=&\rho(h_1,h_2)\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}.
\end{aligned}
$$

□

C.2 Deviation Bounds

In this section, we demonstrate deviation bounds for our error estimators on S_k. Again, unless otherwise specified, all probabilities and expectations in this section are over the random draw of all random variables, that is, $S_0,\cdots,S_K$, $Q_1,\cdots,Q_K$.

We use following Bernstein-style concentration bound:

Fact C.9. *Suppose $X_1,\ldots,X_n$ are independent random variables. For any $i=1,\ldots,n$, $|X_i|\leq 1$,*

$\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 \le \sigma_i^2$. *Then with probability at least* $1-\delta$,

$$\left|\sum_{i=1}^n X_i\right| \le \frac{2}{3}\log\frac{2}{\delta} + \sqrt{2\sum_{i=1}^n \sigma_i^2 \log\frac{2}{\delta}}.$$

Theorem C.10. *For any* $k = 0,\dots,K$, *any* $\delta > 0$, *with probability at least* $1-\delta$, *for all* $h_1, h_2 \in \mathcal{H}$, *the following statement holds:*

$$\begin{aligned}\big|\big(l(h_1,S_k) - l(h_2,S_k)\big) - \big(l(h_1) - l(h_2)\big)\big| \le 2\sup_{x\in X}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\frac{2\log\frac{4|\mathcal{H}|}{\delta}}{3}}{m_kQ_0(x)+n_kQ_k(x)} \\ + \sqrt{2\sup_{x\in X}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}\rho(h_1,h_2)}. \quad \text{(C.1)}\end{aligned}$$

Proof. We show proof for $k > 0$. The $k = 0$ case can be proved similarly. When $k > 0$, it suffices to show that for any $k = 1,\dots,K$, $\delta > 0$, conditioned on Q_k, with probability at least $1-\delta$, (C.1) holds for all $h_1, h_2 \in \mathcal{H}$.

For any $k = 1,\dots,K$, for any fixed $h_1, h_2 \in \mathcal{H}$, define $A := \sup_{x\in X}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}$. Let $N := |S_k|$, $U_t := \frac{\mathbb{1}\{h_1(X_t)\neq Y_t\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)} - \frac{\mathbb{1}\{h_2(X_t)\neq Y_t\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)}$, $V_t := (U_t - \mathbb{E}[U_t|Q_k])/2A$.

Now, conditioned on Q_k, $\{V_t\}_{t=1}^N$ is an independent sequence by Fact C.5. $|V_t| \le 1$, and $\mathbb{E}[V_t|Q_k] = 0$. Besides, we have

$$\begin{aligned}\sum_{t=1}^N \mathbb{E}[V_t^2|Q_k] &\le \frac{1}{4A^2}\sum_{t=1}^N \mathbb{E}[U_t^2|Q_k] \\ &\le \frac{1}{4A^2}\sum_{t=1}^N \mathbb{E}\left(\frac{\mathbb{1}\{h_1(X_t)\neq h_2(X_t)\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)}\right)^2 \\ &\le \frac{\rho(h_1,h_2)}{4A}\end{aligned}$$

where the second inequality follows from $|U_t| \leq \frac{\mathbb{1}\{h_1(X_t) \neq h_2(X_t)\} Z_t}{m_k Q_0(X_t) + n_k Q_k(X_t)}$, and the third inequality follows from Lemma C.8.

Applying Bernstein's inequality (Fact C.9) to $\{V_t\}$, conditioned on Q_k, we have with probability at least $1 - \delta$,

$$\left| \sum_{t=1}^{m} V_t \right| \leq \frac{2}{3} \log \frac{2}{\delta} + \sqrt{\frac{\rho(h_1, h_2)}{2A} \log \frac{2}{\delta}}.$$

Now, $\sum_{t=1}^{m} U_t = l(h_1, S_k) - l(h_2, S_k)$, and $\sum_{t=1}^{m} \mathbb{E}[U_t \mid Q_k] = l(h_1) - l(h_2)$ by Lemma C.7, so $\sum_{t=1}^{m} V_t = \frac{1}{2A}(l(h_1, S_k) - l(h_2, S_k) - l(h_1) + l(h_2))$. (C.1) follows by algebra and a union bound over $\mathcal{H}$. □

Theorem C.11. *For any* $k = 0, \ldots, K$, *any* $\delta > 0$, *with probability at least* $1 - \delta$, *for all* $h_1, h_2 \in \mathcal{H}$, *the following statements hold simultaneously:*

$$\rho_{S_k}(h_1, h_2) \leq 2\rho(h_1, h_2) + \frac{10}{3} \sup_{x \in X} \frac{\mathbb{1}\{h_1(x) \neq h_2(x)\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)}; \tag{C.2}$$

$$\rho(h_1, h_2) \leq 2\rho_{S_k}(h_1, h_2) + \frac{7}{6} \sup_{x \in X} \frac{\mathbb{1}\{h_1(x) \neq h_2(x)\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)}. \tag{C.3}$$

Proof. Let $N = |S_k|$. Note that for any $h_1, h_2 \in \mathcal{H}$, $\rho_{S_k}(h_1, h_2) = \frac{1}{N} \sum_t \mathbb{1}\{h_1(X_t) \neq h_2(X_t)\}$, which is the empirical average of an i.i.d. sequence. By Fact C.9 and a union bound over $\mathcal{H}$, with probability at least $1 - \delta$,

$$\left| \rho(h_1, h_2) - \rho_{S_k}(h_1, h_2) \right| \leq \frac{2}{3N} \log \frac{4|\mathcal{H}|}{\delta} + \sqrt{\frac{2\rho(h_1, h_2)}{N} \log \frac{4|\mathcal{H}|}{\delta}}.$$

On this event, by Proposition C.1, $\rho(h_1, h_2) \leq 2\rho_{S_k}(h_1, h_2) + \frac{4}{3N} \log \frac{4|\mathcal{H}|}{\delta} + \frac{2}{N} \log \frac{4|\mathcal{H}|}{\delta} \leq 2\rho_{S_k}(h_1, h_2) + \frac{10}{3N} \log \frac{4|\mathcal{H}|}{\delta}$.

Moreover,

$$\begin{aligned}
\rho_{S_k}(h_1,h_2) &\leq \rho(h_1,h_2)+\frac{2}{3N}\log\frac{4|\mathcal{H}|}{\delta}+\sqrt{\frac{2\rho(h_1,h_2)}{N}\log\frac{4|\mathcal{H}|}{\delta}} \\
&\leq \rho(h_1,h_2)+\frac{2}{3N}\log\frac{4|\mathcal{H}|}{\delta}+\frac{1}{2}(2\rho(h_1,h_2)+\frac{1}{N}\log\frac{4|\mathcal{H}|}{\delta}) \\
&\leq 2\rho(h_1,h_2)+\frac{7}{6N}\log\frac{4|\mathcal{H}|}{\delta}
\end{aligned}$$

where the second inequality uses the fact that $\forall a,b>0, \sqrt{ab}\leq\frac{a+b}{2}$.

The result follows by noting that $\forall x\in\mathcal{X}, N=|S_k|=m_k+n_k\geq m_kQ_0(x)+n_kQ_k(x)$. □

Corollary C.12. *There are universal constants $\gamma_0,\gamma_1>0$ such that for any $k=0,\ldots,K$, any $\delta>0$, with probability at least $1-\delta$, for all $h,h_1,h_2\in\mathcal{H}$, the following statements hold simultaneously:*

$$\begin{aligned}
\left|\big(l(h_1,S_k)-l(h_2,S_k)\big)-\big(l(h_1)-l(h_2)\big)\right| \leq &\gamma_0\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{|\mathcal{H}|}{2\delta}}{m_kQ_0(x)+n_kQ_k(x)} \\
&+\gamma_0\sqrt{\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{|\mathcal{H}|}{2\delta}}{m_kQ_0(x)+n_kQ_k(x)}\rho_S(h_1,h_2)};
\end{aligned} \tag{C.4}$$

$$\begin{aligned}
l(h)-l(h^\star)\leq &2(l(h,S_k)-l(h^\star,S_k))+\gamma_1\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\log\frac{|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)} \\
&+\gamma_1\sqrt{\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\log\frac{|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}l(h^\star)}.
\end{aligned} \tag{C.5}$$

Proof. Let event E be the event that (C.1) and (C.3) holds for all $h_1,h_2\in\mathcal{H}$ with confidence $1-\frac{\delta}{2}$ respectively. Assume E happens (whose probability is at least $1-\delta$).

(C.4) is immediate from (C.1) and (C.3).

For the proof of (C.5), apply (C.1) to h and $h^\star$, we get

$$\begin{aligned} l(h)-l(h^\star) \le & l(h,S_k)-l(h^\star,S_k)+2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\frac{2\log\frac{4|\mathcal{H}|}{\delta}}{3}}{m_kQ_0(x)+n_kQ_k(x)} \\ & +\sqrt{2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}\rho(h,h^\star)}. \end{aligned}$$

By triangle inequality, $\rho(h,h^\star)=\mathbb{P}_D(h(X)\neq h^\star(X))\le\mathbb{P}_D(h(X)\neq Y)+\mathbb{P}_D(h^\star(X)\neq Y)=l(h)-l(h^\star)+2l(h^\star)$. Therefore, we get

$$\begin{aligned} l(h)-l(h^\star) &\le l(h,S_k)-l(h^\star,S_k)+2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\frac{2\log\frac{4|\mathcal{H}|}{\delta}}{3}}{m_kQ_0(x)+n_kQ_k(x)} \\ &\quad +\sqrt{2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x))\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}(l(h)-l(h^\star)+2l(h^\star))} \\ &\le l(h,S_k)-l(h^\star,S_k)+\sqrt{2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}(l(h)-l(h^\star))} \\ &\quad +2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\frac{2\log\frac{4|\mathcal{H}|}{\delta}}{3}}{m_kQ_0(x)+n_kQ_k(x)}+\sqrt{4\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h(x)\neq h^\star(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}l(h^\star)} \end{aligned}$$

where the second inequality uses $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$.

(C.5) follows by applying Proposition C.1 to $l(h)-l(h^\star)$. □

C.3 Technical Lemmas

For any $0\le k\le K$ and $\delta>0$, define event $\mathcal{E}_{k,\delta}$ to be the event that the conclusions of Theorem C.10 and Theorem C.11 hold for k with confidence $1-\delta/2$ respectively. We have

$\mathbb{P}(\mathcal{E}_{k,\delta}) \geq 1-\delta$, and that $\mathcal{E}_{k,\delta}$ implies inequalities (C.1) to (C.5).

We first present a lemma which can be used to guarantee that $h^\star$ stays in candidate sets with high probability by induction..

Lemma C.13. *For any $k = 0, \ldots K$, any $\delta > 0$. On event $\mathcal{E}_{k,\delta}$, if $h^\star \in C_k$ then,*

$$l(h^\star, \tilde{S}_k) \leq l(\hat{h}_k, \tilde{S}_k) + \gamma_0 \sigma(k,\delta) + \gamma_0 \sqrt{\sigma(k,\delta) \rho_{\tilde{S}_k}(\hat{h}_k, h^\star)}.$$

Proof.

$$\begin{aligned}
&l(h^\star, \tilde{S}_k) - l(\hat{h}_k, \tilde{S}_k) \\
=&l(h^\star, S_k) - l(\hat{h}_k, S_k) \\
\leq&\gamma_0 \sup_x \frac{\mathbb{1}\{h^\star(x) \neq \hat{h}_k(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} + \gamma_0 \sqrt{\sup_x \frac{\mathbb{1}\{h^\star(x) \neq \hat{h}_k(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} \rho_{S_k}(\hat{h}_k, h^\star)} \\
\leq&\gamma_0 \sigma(k,\delta) + \sqrt{\gamma_0 \sigma(k,\delta) \rho_{\tilde{S}_k}(\hat{h}_k, h)}.
\end{aligned}$$

The equality follows from Lemma C.3. The first inequality follows from (C.4) of Corollary C.12 and that $l(h^\star) \leq l(\hat{h}_k)$. The last inequality follows from Lemma C.2 and that $\rho_{\tilde{S}_k}(\hat{h}_k, h^\star) = \rho_{S_k}(\hat{h}_k, h^\star)$. □

Next, we present two lemmas to bound the probability mass of the disagreement region of candidate sets.

Lemma C.14. *For any $k = 0, \ldots, K$, any $\delta > 0$, let $C_{k+1}(\delta) := \{h \in C_k \mid l(h, \tilde{S}_k) \leq l(\hat{h}_k, \tilde{S}_k) + \gamma_0 \sigma(k,\delta) + \gamma_0 \sqrt{\sigma(k,\delta) \rho_{\tilde{S}_k}(\hat{h}_k, h)}\}$. Then there is an absolute constant $\gamma_2 > 1$ such that for any $0, \ldots, K$, any $\delta > 0$, on event $\mathcal{E}_{k,\delta}$, if $h^\star \in C_k$, then for all $h \in C_{k+1}(\delta)$,*

$$l(h) - l(h^\star) \leq \gamma_2 \sigma(k,\delta) + \gamma_2 \sqrt{\sigma(k,\delta) l(h^\star)}.$$

Proof. For any $h \in C_{k+1}(\delta)$, we have

$$
\begin{aligned}
& l(h) - l(h^\star) \\
&\quad \leq 2(l(h,S_k) - l(h^\star,S_k)) + \gamma_1 \sigma(k,\frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \\
&\quad = 2(l(h,\tilde{S}_k) - l(h^\star,\tilde{S}_k)) + \gamma_1 \sigma(k,\frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \\
&\quad = 2(l(h,\tilde{S}_k) - l(\hat{h}_k,\tilde{S}_k) + l(\hat{h}_k,\tilde{S}_k) - l(h^\star,\tilde{S}_k)) + \gamma_1 \sigma(k,\frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \\
&\quad \leq 2(l(h,\tilde{S}_k) - l(\hat{h}_k,\tilde{S}_k)) + \gamma_1 \sigma(k,\frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \\
&\quad \leq (2\gamma_0 + \gamma_1) \sigma(k,\frac{\delta}{2}) + 2\gamma_0 \sqrt{\sigma(k,\frac{\delta}{2}) \rho_{\tilde{S}_k}(h,\hat{h}_k)} + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \\
&\quad \leq (2\gamma_0 + \gamma_1) \sigma(k,\frac{\delta}{2}) + 2\gamma_0 \sqrt{\sigma(k,\frac{\delta}{2}) (\rho_{S_k}(h,h^\star) + \rho_{S_k}(\hat{h}_k,h^\star))} + \gamma_1 \sqrt{\sigma(k,\frac{\delta}{2}) l(h^\star)} \qquad \text{(C.6)}
\end{aligned}
$$

where the first inequality follows from (C.5) of Corollary C.12 and Lemma C.2, the first equality follows from Lemma C.3, the third inequality follows from the definition of $C_k(\delta)$, and the last inequality follows from $\rho_{\tilde{S}_k}(h,\hat{h}_k) = \rho_{S_k}(h,\hat{h}_k) \leq \rho_{S_k}(h,h^\star) + \rho_{S_k}(\hat{h}_k,h^\star)$.

As for $\rho_{S_k}(h,h^\star)$, we have $\rho_{S_k}(h,h^\star) \leq 2\rho(h,h^\star) + \frac{16}{3}\sigma(k,\frac{\delta}{8}) \leq 2(l(h) - l(h^\star)) + 4l(h^\star) + \frac{16}{3}\sigma(k,\frac{\delta}{8})$ where the first inequality follows from (C.2) of Theorem C.11 and Lemma C.2, and the second inequality follows from the triangle inequality.

For $\rho_{S_k}(\hat{h}_k, h^\star)$, we have

$$\begin{aligned}
\rho_{S_k}(\hat{h}_k, h^\star) &\leq 2\rho(\hat{h}_k, h^\star) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq 2(l(\hat{h}_k) - l(h^\star) + 2l(h^\star)) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq 2(2(l(\hat{h}_k, S_k) - l(h^\star, S_k)) + \gamma_1\sigma(k, \frac{\delta}{2}) + \gamma_1\sqrt{\sigma(k, \frac{\delta}{2})l(h^\star)} + 2l(h^\star)) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq (2\gamma_1 + \frac{16}{3})\sigma(k, \frac{\delta}{8}) + 2\gamma_1\sqrt{\sigma(k, \frac{\delta}{2})l(h^\star)} + 4l(h^\star) \\
&\leq (4 + \gamma_1)l(h^\star) + (3\gamma_1 + \frac{16}{3})\sigma(k, \frac{\delta}{8})
\end{aligned}$$

where the first inequality follows from (C.2) of Theorem C.11 and Lemma C.2, the second follows from the triangle inequality, the third follows from (C.5) of Theorem C.12 and Lemma C.2, the fourth follows from the definition of $\hat{h}_k$, the last follows from the fact that $2\sqrt{ab} \leq a + b$ for $a, b \geq 0$.

Continuing (C.6) and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have:

$$\begin{aligned}
l(h) - l(h^\star) \leq &(2\gamma_0 + \gamma_1 + 2\gamma_0\sqrt{3\gamma_1 + \frac{32}{3}})\sigma(k, \frac{\delta}{8}) \\
&+ (2\gamma_0\sqrt{8 + \gamma_1} + \gamma_1)\sqrt{\sigma(k, \frac{\delta}{8})l(h^\star)} + 2\sqrt{2}\gamma_0\sqrt{\sigma(k, \frac{\delta}{8})(l(h) - l(h^\star))}.
\end{aligned}$$

The result follows by applying Proposition C.1 to $l(h) - l(h^\star)$. □

Lemma C.15. *On event $\bigcap_{k=0}^{K-1} \mathcal{E}_{k, \delta_k/2}$, for any $k = 0, \ldots K$, $D_k \subseteq DIS_k$.*

Proof. Recall $\delta_k = \frac{\delta}{(k+1)(k+2)}$. On event $\bigcap_{k=0}^{K-1} \mathcal{E}_{k, \delta_k/2}$, $h^\star \in C_k$ for all $0 \leq k \leq K$ by Lemma C.13 and induction.

The $k = 0$ case is obvious since $D_0 = \text{DIS}_0 = \mathcal{X}$. Now, suppose $0 \leq k < K$, and $D_k \subseteq \text{DIS}_k$.

We have

$$\begin{aligned} D_{k+1} &\subseteq \mathrm{DIS}\left(\left\{h : l(h) \le \nu + \gamma_2\left(\sigma(k,\delta_k/2) + \sqrt{\sigma(k,\delta_k/2)\nu}\right)\right\}\right) \\ &\subseteq \mathrm{DIS}\left(B\left(h^\star, 2\nu + \gamma_2\left(\sigma(k,\delta_k/2) + \sqrt{\sigma(k,\delta_k/2)\nu}\right)\right)\right) \end{aligned}$$

where the first line follows from Lemma C.14 and the definition of D_k, and the second line follows from triangle inequality that $\mathbb{P}(h(X) \neq h^\star(X)) \le l(h) + l(h^\star)$ (recall $\nu = l(h^\star)$).

To prove $D_{k+1} \subseteq \mathrm{DIS}_{k+1}$ it suffices to show $\gamma_2\left(\sigma(k,\delta_k/2) + \sqrt{\sigma(k,\delta_k/2)\nu}\right) \le \varepsilon_{k+1}$.

Note that $\sigma(k,\delta_k/2) = \sup_{x \in D_k} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_k Q_0(x) + n_k} \le \sup_{x \in \mathrm{DIS}_k} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_k Q_0(x) + n_k}$ since $D_k \subseteq \mathrm{DIS}_k$. Consequently, $\gamma_2\left(\sigma(k,\delta_k/2) + \sqrt{\sigma(k,\delta_k/2)\nu}\right) \le \varepsilon_{k+1}$. □

C.4 Proof of Consistency

Proof. (of Theorem 6.1) Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K} \mathcal{E}_{k,\delta_k/2}$. By a union bound, $\mathbb{P}(\mathcal{E}^{(0)}) \ge 1 - \delta$. On event $\mathcal{E}^{(0)}$, by induction and Lemma C.13, for all $k = 0, \ldots, K$, $h^\star \in C_k$. Observe that $\hat{h} = \hat{h}_K \in C_{K+1}(\delta_K/2)$. Applying Lemma C.14 to $\hat{h}$, we have

$$l(\hat{h}) \le l(h^\star) + \gamma_2\left(\sup_{x \in D_K} \frac{\log(2|\mathcal{H}|/\delta_K)}{m_K Q_0(x) + n_K} + \sqrt{\sup_{x \in D_K} \frac{\log(2|\mathcal{H}|/\delta_K)}{m_K Q_0(x) + n_K} l(h^\star)}\right).$$

The result follows by noting that $\sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{x \in D_K\}}{m_K Q_0(x) + n_K} \le \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{x \in \mathrm{DIS}_K\}}{m_K Q_0(x) + n_K}$ by Lemma C.15. □

C.5 Proof of Label Complexity

Proof. (of Theorem 6.4) Recall that $\xi_k = \inf_{x\in D_k} Q_0(x)$ and $\zeta = \sup_{x\in \mathrm{DIS}_1} \frac{1}{\alpha Q_0(x)+1}$.

Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K} \mathcal{E}_{k,\delta_k/2}$. On this event, by induction and Lemma C.13, for all $k = 0,\dots,K$, $h^\star \in C_k$, and consequently by Lemma C.15, $D_k \subseteq \mathrm{DIS}_k$.

For any $k = 0,\dots K-1$, let the number of label queries at iteration k to be $U_k := \sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} Z_t \mathbb{1}\{X_t \in D_{k+1}\}$.

$$\begin{aligned} Z_t \mathbb{1}\{X_t \in D_{k+1}\} &= \mathbb{1}\{X_t \in D_{k+1} \wedge Q_0(X_t) \le \inf_{x\in D_{k+1}} Q_0(x) + \frac{1}{\alpha}\} \\ &\le \mathbb{1}\{X_t \in S(D_{k+1},\alpha)\} \\ &\le \mathbb{1}\{X_t \in S(\mathrm{DIS}_{k+1},\alpha)\}. \end{aligned}$$

Thus, $U_k \le \sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} \mathbb{1}\{X_t \in S(\mathrm{DIS}_{k+1},\alpha)\}$, where the RHS is sum of i.i.d. Bernoulli random variables with mean $\mathbb{P}(S(\mathrm{DIS}_{k+1},\alpha))$, so a Bernstein inequality implies that on an event $\mathcal{E}^{(1,k)}$ of probability at least $1-\delta_k/2$,

$$\sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} \mathbb{1}\{X_t \in S(\mathrm{DIS}_{k+1},\alpha)\} \le 2n_{k+1}\mathbb{P}(S(\mathrm{DIS}_{k+1},\alpha)) + 2\log\frac{4}{\delta_k}$$

.

Therefore, it suffices to show that on event $\mathcal{E}^{(2)} := \cap_{k=0}^{K}(\mathcal{E}^{(1,k)} \cap \mathcal{E}_{k,\delta_k/2})$, for some absolute constant c_1, $\sum_{k=0}^{K-1} n_{k+1}\mathbb{P}(S(\mathrm{DIS}_{k+1},\alpha))$ is at most

$$c_1\tilde{\theta}(2\nu+\varepsilon_K,\alpha)(n\nu + \zeta\log n\log\frac{|\mathcal{H}|\log n}{\delta} + \log n\sqrt{n\nu\zeta\log\frac{|\mathcal{H}|\log n}{\delta}}).$$

Now, on event $\mathcal{E}^{(2)}$, for any $k < K$, $\mathbb{P}(S(\text{DIS}_{k+1}, \alpha)) = \mathbb{P}(S(\text{DIS}(B(h^\star, 2\nu + \varepsilon_{k+1})), \alpha)) \leq (2\nu + \varepsilon_{k+1})\tilde{\theta}(2\nu + \varepsilon_{k+1}, \alpha)$ where the last inequality follows from Lemma C.4.

Therefore,

$$\begin{aligned}
&\sum_{k=0}^{K-1} n_{k+1}\mathbb{P}(S(\text{DIS}_{k+1}, \alpha)) \\
&\quad\leq n_1 + \sum_{k=1}^{K-1} n_{k+1}(2\nu + \varepsilon_{k+1})\tilde{\theta}(2\nu + \varepsilon_{k+1}, \alpha) \\
&\quad\leq 1 + \tilde{\theta}(2\nu + \varepsilon_K, \alpha)(2n\nu + \sum_{k=1}^{K-1} n_{k+1}\varepsilon_{k+1}) \\
&\quad\leq 1 + \tilde{\theta}(2\nu + \varepsilon_K, \alpha)\left(2n\nu + 2\gamma_2 \sum_{k=1}^{K-1} (\sup_{x \in \text{DIS}_1} \frac{\log \frac{|\mathcal{H}|}{\delta_k/2}}{(\alpha Q_0(x) + 1)} + \sqrt{n_k \nu \sup_{x \in \text{DIS}_1} \frac{\log \frac{|\mathcal{H}|}{\delta_k/2}}{(\alpha Q_0(x) + 1)}})\right) \\
&\quad\leq 1 + \tilde{\theta}(2\nu + \varepsilon_K, \alpha)(2n\nu + 2\gamma_2 \zeta \log n \log \frac{|\mathcal{H}|(\log n)^2}{\delta} + 2\gamma_2 \log n \sqrt{n\nu\zeta \log \frac{|\mathcal{H}|(\log n)^2}{\delta}}).
\end{aligned}$$

□

Appendix D

Omitted Proofs for Chapter 7

D.1 Preliminaries

D.1.1 Summary of Key Notations

Data $T_0 = \{(X_t, Y_t, Z_t)\}_{t=1}^{m}$ is the logged data. $\tilde{T}_k = \{(X_t, \tilde{Y}_t, Z_t)\}_{t=m+n_{k-1}+1}^{m+n_k}$ $(1 \leq k \leq K)$ is the online data collected in the k-th iteration of size $\tau_k = n_k - n_{k-1}$, and $\tilde{Y}_t$ equals either the actual label Y_t drawn from the data distribution D or the inferred label $\hat{h}_{k-1}(X_t)$ according to the candidate set C_{k-1}at iteration $k-1$. $\tilde{S}_k = T_0 \cup \tilde{T}_1 \cup \cdots \cup \tilde{T}_k$.

For convenience, we additionally define $T_k = \{(X_t, Y_t, Z_t)\}_{t=m+n_{k-1}+1}^{m+n_k}$ to be the data set with the actual labels Y_t drawn from the data distribution, and $S_k = T_0 \cup T_1 \cup \cdots \cup T_k$. The algorithm only observes $\tilde{S}_k$ and $\tilde{T}_k$, and S_k, T_k are used for analysis only.

For $1 \leq k \leq K$,$n_k = \tau_1 + \cdots + \tau_k$, and we define $n_0 = 0$, $n = n_K$, $\tau_0 = m$. We assume $\tau_k \leq \tau_{k+1}$ for $1 \leq k < K$.

Recall that $\{(X_t,Y_t,Z_t)\}_{t=1}^{m+n}$ is an independent sequence, and furthermore $\{(X_t,Y_t)\}_{t=1}^{m+n}$ is an i.i.d. sequence drawn from D. For $(X,Z) \in T_k$ $(0 \leq k \leq K)$, $Q_k(X) = \mathbb{P}(Z = 1 \mid X)$. Unless otherwise specified, all probabilities and expectations are over the random draw of all random variables $\{(X_t,Y_t,Z_t)\}_{t=1}^{m+n}$.

Loss and Second Moment The test error $l(h) = \mathbb{P}(h(X) \neq Y)$, the optimal classifier $h^\star = \arg\min_{h \in \mathcal{H}} l(h)$, and the optimal error $\nu = l(h^\star)$. At the k-th iteration, the Multiple Importance Sampling (MIS) weight $w_k(x) = \frac{m+n_k}{mQ_0(X_t)+\sum_{i=1}^{k}\tau_i Q_i(X_t)}$. The clipped MIS loss estimator $l(h;S_k,M) = \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k(X_i) Z_i \mathbb{1}\{h(X_i) \neq Y_i\}\mathbb{1}\{w_k(X_i) \leq M\}$. The (unclipped) MIS loss estimator $l(h;S_k) = l(h;S_k,\infty)$.

The clipped second moment $\mathrm{V}(h;k,M) = \mathbb{E}\left[w_k(X)\mathbb{1}\{h(X) \neq Y\}\mathbb{1}\{w_k(X) \leq M\}\right]$, and $\mathrm{V}(h_1,h_2;k,M) = \mathbb{E}\left[w_k(X)\mathbb{1}\{h_1(X) \neq h_2(X)\}\mathbb{1}\{w_k(X) \leq M\}\right]$.

The clipped second-moment estimators $\hat{\mathrm{V}}(h;S_k,M) = \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k^2(X_i) Z_i \mathbb{1}\{h(X_i) \neq Y_i\}\mathbb{1}\{w_k(X_i) \leq M\}$, $\hat{\mathrm{V}}(h_1,h_2;S_k,M) = \frac{1}{m+n_k}\sum_{i=1}^{m+n_k} w_k^2(X_i) Z_i \mathbb{1}\{h_1(X) \neq h_2(X)\}\mathbb{1}\{w_k(X_i) \leq M\}$.

The unclipped second moments ($\mathrm{V}(h;k)$, $\mathrm{V}(h_1,h_2;k)$) and second moment estimators ($\hat{\mathrm{V}}(h;S_k)$, $\hat{\mathrm{V}}(h_1,h_2;S_k)$) are defined similarly.

Disagreement Regions The r-ball around h is defined as $B(h,r) := \{h' \in \mathcal{H} \mid \mathbb{P}(h(X) \neq h'(X)) \leq r\}$. The disagreement region of $C \subseteq \mathcal{H}$ is $\mathrm{DIS}(C) := \{x \in \mathcal{X} \mid \exists h_1 \neq h_2 \in C \text{ s.t. } h_1(x) \neq h_2(x)\}$.

The candidate set C_k and its disagreement region D_k are defined in Algorithm 11. The empirical risk minimizer (ERM) at k-th iteration $\hat{h}_k = \arg\min_{h \in C_k} l(h,\tilde{S}_k)$.

The modified disagreement coefficient $\tilde{\theta}(r,\alpha) := \frac{1}{r}\mathbb{P}\left(\mathrm{DIS}(B(h^\star,r)) \cap \left\{x : Q_0(x) \leq \frac{1}{\alpha}\right\}\right)$.

$\tilde{\theta} = \sup_{r>2\nu} \tilde{\theta}(r, \frac{2m}{n})$.

Other Notations $q_0 = \inf_x Q_0(x)$. $Q_{k+1}(x) = \mathbb{1}\{mQ_0(x) + \sum_{i=1}^{k} \tau_i Q_i(x) < \frac{m}{2} Q_0(x) + n_{k+1}\}$. $M_k = \inf\{M \geq 1 \mid \frac{2M}{m+n_k} \log \frac{|\mathcal{H}|}{\delta_k} \geq \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} > M/2)\}$. $\xi = \min_{1 \leq k \leq K}\{M_k / \frac{m+n_k}{mq_0+n_k}\}$. $\bar{M} = \max_{1 \leq k \leq K} M_k$.

D.1.2 Elementary Facts

Proposition D.1. *Suppose $a, c \geq 0$, $b \in \mathbb{R}$. If $a \leq b + \sqrt{ca}$, then $a \leq 2b + c$.*

Proof. Since $a \leq b + \sqrt{ca}$, $\sqrt{a} \leq \frac{\sqrt{c}+\sqrt{c+4b}}{2} \leq \sqrt{\frac{c+c+4b}{2}} = \sqrt{c+2b}$ where the second inequality follows from the Root-Mean Square-Arithmetic Mean inequality. Thus, $a \leq 2b + c$. □

D.1.3 Facts on Disagreement Regions and Candidate Sets

Lemma D.2. *For any $k = 0, \ldots, K$, $M \geq 0$, if $h_1, h_2 \in C_k$, then $l(h_1; S_k, M) - l(h_2; S_k, M) = l(h_1; \tilde{S}_k, M) - l(h_2; \tilde{S}_k, M)$ and $\hat{V}(h_1, h_2; S_k, M) = \hat{V}(h_1, h_2; \tilde{S}_k, M)$.*

Proof. For any $(X_t, Y_t, Z_t) \in S_k$ that $Z_t = 1$, if $X_t \in \text{DIS}(C_k)$, then $Y_t = \tilde{Y}_t$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\}$. If $X_t \notin \text{DIS}(C_k)$, then $h_1(X_t) = h_2(X_t)$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\} = 0$. Thus, $l(h_1; S_k, M) - l(h_2; S_k, M) = l(h_1; \tilde{S}_k, M) - l(h_2; \tilde{S}_k, M)$.

$\hat{V}(h_1, h_2; S_k, M) = \hat{V}(h_1, h_2; \tilde{S}_k, M)$ holds since $\hat{V}(h_1, h_2; S_k, M)$ and $\hat{V}(h_1, h_2; \tilde{S}_k, M)$ do not involve labels Y or $\tilde{Y}$. □

The following lemmas are immediate from the definition.

Lemma D.3. *For any* $1 \leq k \leq K$*, if* $h \in C_k$*, then* $l(h;\tilde{S}_k,M) \leq l(h;S_k,M) \leq l(h;S_k)$ *and* $\hat{V}(h;\tilde{S}_k,M) \leq \hat{V}(h;S_k,M) \leq \hat{V}(h;S_k)$*.*

Remark D.4. The inequality on the second moment regularizer $\hat{\mathrm{V}}$, which will be used to prove the error bound (Theorem 7.9) of Algorithm 11, is due to the decomposition property $\hat{\mathrm{V}}(h;S_k,M) = \frac{|S_k \cap \mathrm{DIS}(C_k)|}{m+n_k}\hat{\mathrm{V}}(h;S_k \cap \mathrm{DIS}(C_k),M) + \frac{|S_k \cap \mathrm{DIS}(C_k)^c|}{m+n_k}\hat{\mathrm{V}}(h;S_k \cap \mathrm{DIS}(C_k)^c,M)$. It does not hold for estimated variance $\hat{\mathrm{Var}}(h;S_k,M) := \hat{\mathrm{V}}(h;S_k,M) - l(h;S_k,M)^2$. This explains the necessity of introducing the second moment regularizer.

Lemma D.5. *For any* $r \geq 2\nu$*, any* $\alpha \geq 1$*,* $\mathbb{P}(DIS(B(h^\star,r) \cap \{x : Q_0(x) \leq \frac{1}{\alpha}\}) \leq r\tilde{\theta}(r,\alpha)$*.*

D.1.4 Facts on Multiple Importance Sampling Estimators

Proposition D.6. *Let* $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$*. For any k, the following equations hold:*

$$\mathbb{E}[\frac{1}{m+n_k} \sum_{(X,Y,Z) \in S_k} w_k(X) Z f(X,Y)] = \mathbb{E}[f(X,Y)],$$
$$\mathbb{E}[\frac{1}{m+n_k} \sum_{(X,Y,Z) \in S_k} w_k^2(X) Z f(X,Y)] = \mathbb{E}[w_k(X) f(X,Y)].$$

Proof.

$$\begin{aligned}
\mathbb{E}[\sum_{(X,Y,Z)\in S_k} w_k(X)Zf(X,Y)] &= \sum_{i=0}^{k}\mathbb{E}[\sum_{(X,Y,Z)\in T_i}\mathbb{E}[w_k(X)f(X,Y)Z \mid X,Y]] \\
&= \sum_{i=0}^{k}\mathbb{E}[\sum_{(X,Y,Z)\in T_i} w_k(X)f(X,Y)\mathbb{E}[Z \mid X,Y]] \\
&\overset{(a)}{=} \sum_{i=0}^{k}\mathbb{E}[\sum_{(X,Y,Z)\in T_i} w_k(X)f(X,Y)\mathbb{E}[Z \mid X]] \\
&= \sum_{i=0}^{k}\mathbb{E}[\sum_{(X,Y,Z)\in T_i} w_k(X)f(X,Y)Q_i(X)] \\
&\overset{(b)}{=} \sum_{i=0}^{k}\tau_i\mathbb{E}[w_k(X)f(X,Y)Q_i(X)] \\
&= \mathbb{E}[w_k(X)f(X,Y)\sum_{i=0}^{k}\tau_i Q_i(X)] \\
&\overset{(c)}{=} (m+n_k)\mathbb{E}[f(X,Y)]
\end{aligned}$$

where (a) follows from $\mathbb{E}[Z \mid X] = \mathbb{E}[Z \mid X,Y]$ as Z,Y are conditionally independent given X, (b) follows since T_i is a sequence of i.i.d. random variables, and (c) follows from the definition $w_k(X) = \frac{m+n_k}{\sum_{i=0}^{k}\tau_i Q_i(X)}$.

The proof for the second equality is similar and skipped. □

D.1.5 Facts on the Sample Selection Bias Correction Query Strategy

The query strategy Q_k can be simplified as follows.

Proposition D.7. *For any* $1 \le k \le K$, $x \in \mathcal{X}$, $Q_k(x) = \mathbb{1}\{2n_k - mQ_0(x) > 0\}$.

Proof. The $k = 1$ case can be easily verified. Suppose it holds for Q_k, and we next show it holds for Q_{k+1}. Recall by definition $Q_{k+1}(x) = \mathbb{1}\{mQ_0(x) + \sum_{i=1}^{k}\tau_i Q_i(x) < \frac{m}{2}Q_0(x) + n_{k+1}\}$.

If $Q_k(x) = 1$, then $mQ_0(x) + \sum_{i=1}^{k-1} \tau_i Q_i(x) < \frac{m}{2} Q_0(x) + n_k$, so

$$\begin{aligned} mQ_0(x) + \sum_{i=1}^{k} \tau_i Q_i(x) &< \frac{m}{2} Q_0(x) + n_k + \tau_k \\ &\leq \frac{m}{2} Q_0(x) + n_{k+1} \end{aligned}$$

where the last inequality follows by the assumption on the epoch schedule $\tau_k \leq \tau_{k+1} = n_{k+1} - n_k$. This implies $Q_{k+1}(x) = 1$. In this case, $\mathbb{1}\{2n_{k+1} - mQ_0(x) > 0\} = 1$ as well, since $n_{k+1} \geq n_k$ implies $2n_{k+1} - mQ_0(x) \geq 2n_k - mQ_0(x) > 0$.

The above argument also implies if $Q_k(x) = 0$, then $Q_1(x) = Q_2(x) = \cdots = Q_{k-1}(x) = 0$. Thus, if $Q_k(x) = 0$, then $Q_{k+1}(x) = \mathbb{1}\{mQ_0(x) < \frac{m}{2} Q_0(x) + n_{k+1}\} = \mathbb{1}\{2n_{k+1} - mQ_0(x) > 0\}$. □

The following proposition gives an upper bound of the multiple importance sampling weight, which will be used to bound the second moment of the loss estimators with the sample selection bias correction strategy.

Proposition D.8. *For any* $1 \leq k \leq K$, $w_k(x) = \frac{m+n_k}{mQ_0(x) + \sum_{i=1}^{k} \tau_i Q_i(x)} \leq \frac{m+n_k}{\frac{1}{2} mQ_0(x) + n_k}$.

Proof. The $k = 1$ case can be easily verified. Suppose it holds for w_k, and we next show it holds for w_{k+1}.

Now, if $Q_{k+1}(x) = 0$, then by Proposition D.7, $2n_{k+1} - mQ_0(x) \leq 0$, and consequently $mQ_0(x) + \sum_{i=1}^{k+1} \tau_i Q_i(x) \geq mQ_0(x) \geq \frac{1}{2} mQ_0(x) + n_{k+1}$.

If $Q_{k+1}(x) = 1$, then by the induction hypothesis, $mQ_0(x) + \sum_{i=1}^{k+1} \tau_i Q_i(x) \geq \frac{1}{2} mQ_0(x) + n_k + \tau_{k+1} = \frac{1}{2} mQ_0(x) + n_{k+1}$.

In both cases, $mQ_0(x) + \sum_{i=1}^{k+1} \tau_i Q_i(x) \geq \frac{1}{2} mQ_0(x) + n_{k+1}$, so $w_{k+1}(x) \leq \frac{m+n_{k+1}}{\frac{1}{2} mQ_0(x) + n_{k+1}}$. □

D.1.6 Lower Bound Techniques

We present a lower bound for binomial distribution tails, which will be used to prove generalization error lower bounds.

Lemma D.9. *Let $0 < t < p < 1/2$, $B \sim Bin(n,p)$ be a binomial random variable, and $\delta = \sqrt{4n\frac{(t-p)^2}{p}}$. Then, $\mathbb{P}(B < nt) \geq \frac{1}{\sqrt{2\pi}}\frac{\delta}{\delta^2+1}\exp(-\frac{1}{2}\delta^2)$.*

This Lemma is a consequence of following lemmas.

Lemma D.10. *Suppose $0 < p,q < 1$, $KL(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. Then $KL(p,q) \leq \frac{(p-q)^2}{q(1-q)}$.*

Proof. Since $\log x \leq x-1$, $p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q} \leq p(\frac{p}{q}-1)+(1-p)(\frac{1-p}{1-q}-1) = \frac{(p-q)^2}{q(1-q)}$. □

Lemma D.11. *([BS79]) Suppose $X \sim N(0,1)$, and define $\Phi(t) = \mathbb{P}(X \leq t)$. If $t > 0$, then $\Phi(-t) \geq \frac{1}{\sqrt{2\pi}}\frac{t}{t^2+1}\exp(-\frac{1}{2}t^2)$.*

Lemma D.12. *([ZS13]) Let $B \sim Bin(n,p)$ be a binomial random variable and $0 < k < np$. Then, $\mathbb{P}(B < k) \geq \Phi(-\sqrt{2nKL(\frac{k}{n},p)})$.*

D.2 Deviation Bounds

In this section, we demonstrate deviation bounds for our error estimators on S_k.

We use following Bernstein-style concentration bound:

Fact D.13. *Suppose $X_1,\ldots,X_n$ are independent random variables such that $|X_i| \leq M$. Then with* *probability at least* $1-\delta$*,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i\right| \le \frac{2M}{3n}\log\frac{2}{\delta} + \sqrt{\frac{2}{n^2}\sum_{i=1}^{n}\mathbb{E}X_i^2\log\frac{2}{\delta}}.$$

Theorem D.14. *For any* $k=0,\ldots,K$*, any* $\delta>0$*, if* $\frac{2M\log\frac{|\mathcal{H}|}{\delta}}{m+n_k} \ge \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} \ge \frac{M}{2})$*, then with probability at least* $1-\delta$*, for all* $h_1,h_2\in\mathcal{H}$*, the following statements hold simultaneously:*

$$\left|\big(l(h_1;S_k,M)-l(h_2;S_k,M)\big)-\big(l(h_1)-l(h_2)\big)\right| \le \frac{10\log\frac{2|\mathcal{H}|}{\delta}}{3(m+n_k)}M + \sqrt{\frac{4\log\frac{2|\mathcal{H}|}{\delta}}{m+n_k}V(h_1,h_2;k,M)}; \tag{D.1}$$

$$\left|l(h_1;S_k,M)-l(h_1)\right| \le \frac{10\log\frac{2|\mathcal{H}|}{\delta}}{3(m+n_k)}M + \sqrt{\frac{4\log\frac{2|\mathcal{H}|}{\delta}}{m+n_k}V(h_1;k,M)}. \tag{D.2}$$

Proof. We show proof for $k>0$. The $k=0$ case can be proved similarly.

First, define the clipped expected loss $l(h;k,M)=\mathbb{E}[\mathbb{1}\{h(X)\neq Y\}\mathbb{1}\{w_k(X)\le M\}]$. We have

$$\begin{aligned}
&\left|\big(l(h_1)-l(h_2)\big)-\big(l(h_1;k,M)-l(h_2;k,M)\big)\right| \\
=&\left|\mathbb{E}\left[(\mathbb{1}\{h_1(X)\neq Y\}-\mathbb{1}\{h_2(X)\neq Y\})\mathbb{1}\{w_k(X)>M\}\right]\right| \\
\le&\mathbb{E}\left[\mathbb{1}[w_k(X)>M]\right] \\
\le&\mathbb{E}[\mathbb{1}\{\frac{m+n_k}{mQ_0(X)+n_k}>\frac{M}{2}\}] \\
\le&\frac{2M}{m+n_k}\log\frac{|\mathcal{H}|}{\delta}
\end{aligned} \tag{D.3}$$

where the second inequality follows from Proposition D.8, and the last inequality follows from the assumption on M.

Next, we bound $\big(l(h_1;S_k,M) - l(h_2;S_k,M)\big) - \big(l(h_1;k,M) - l(h_2;k,M)\big)$.

For any fixed $h_1, h_2 \in \mathcal{H}$, define $N := |S_k|$, $U_t := w_k(X_t) Z_t \mathbb{1}\{w_k(X_t) \leq M\}(\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\})$.

Now, $\{U_t\}_{t=1}^N$ is an independent sequence. $\frac{1}{N}\sum_{t=1}^N U_t = l(h_1;S_k,M) - l(h_2;S_k,M)$, and $\mathbb{E}\frac{1}{N}\sum_{t=1}^N U_t = l(h_1;k,M) - l(h_2;k,M)$ by Proposition D.6. Moreover, since $(\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\})^2 = \mathbb{1}\{h_1(X_t) \neq h_2(X_t)\}$, we have $\frac{1}{N}\sum_{t=1}^N U_t^2 = \hat{\mathrm{V}}(h_1,h_2;S_k,M)$ and by Proposition D.6 $\mathbb{E}\frac{1}{N}\sum_{t=1}^N U_t^2 = \mathrm{V}(h_1,h_2;k,M)$. Applying Bernstein's inequality (Fact D.13) to $\{U_t\}$, we have with probability at least $1 - \frac{\delta}{2}$,

$$\left|\frac{1}{N}\sum_{t=1}^N U_t - \mathbb{E}\frac{1}{N}\sum_{t=1}^N U_t\right| \leq \frac{2M}{3N}\log\frac{4}{\delta} + \sqrt{\frac{2}{N}\mathrm{V}(h_1,h_2;k,M)\log\frac{4}{\delta}},$$

and consequently $|\big(l(h_1;S_k,M) - l(h_2;S_k,M)\big) - \big(l(h_1;k,M) - l(h_2;k,M)\big)| \leq \frac{2M}{3(m+n_k)}\log\frac{4}{\delta} + \sqrt{\frac{2}{m+n_k}\mathrm{V}(h_1,h_2;k,M)\log\frac{4}{\delta}}$.

By a union bound over $\mathcal{H}$, with probability at least $1 - \frac{\delta}{2}$ for all $h_1, h_2 \in \mathcal{H}$,

$$\begin{aligned}&\left|\big(l(h_1;S_k,M) - l(h_2;S_k,M)\big) - \big(l(h_1;k,M) - l(h_2;k,M)\big)\right| \\ &\leq \frac{4M}{3(m+n_k)}\log\frac{2|\mathcal{H}|}{\delta} + \sqrt{\frac{4}{m+n_k}\mathrm{V}(h_1,h_2;k,M)\log\frac{2|\mathcal{H}|}{\delta}}.\end{aligned} \tag{D.4}$$

(D.1) follows by combining (D.3) and (D.4).

The proof for (D.2) is similar and skipped. □

We use following bound for the second moment which is an immediate corollary of

Lemmas B.1 and B.2 in [ND17]:

Fact D.15. *Suppose $X_1,\dots,X_n$ are independent random variables such that $|X_i| \leq M$. Then with probability at least $1-\delta$,*

$$-\sqrt{\frac{2M^2}{n}\log\frac{1}{\delta}}-\frac{M^2}{n} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}-\sqrt{\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}X_i^2} \leq \sqrt{\frac{2M^2}{n}\log\frac{1}{\delta}}.$$

Recall that by Lemma D.6, $\mathbb{E}[\hat{\mathrm{V}}(h_1,h_2;S_k,M)] = \mathrm{V}(h_1,h_2;k,M)$ and $\mathbb{E}[\hat{\mathrm{V}}(h_1;S_k,M)] = \mathrm{V}(h_1;k,M)$. The following Corollary follows from the bound on the second moment.

Corollary D.16. *For any $k=0,\dots,K$, any $\delta, M > 0$, with probability at least $1-\delta$, for all $h_1,h_2 \in \mathcal{H}$, the following statements hold:*

$$\left|\sqrt{\hat{\mathrm{V}}(h_1,h_2;S_k,M)}-\sqrt{\mathrm{V}(h_1,h_2;k,M)}\right| \leq \sqrt{\frac{4M^2}{m+n_k}\log\frac{2|\mathcal{H}|}{\delta}}+\frac{M^2}{m+n_k}, \qquad \text{(D.5)}$$

$$\left|\sqrt{\hat{\mathrm{V}}(h_1;S_k,M)}-\sqrt{\mathrm{V}(h_1;k,M)}\right| \leq \sqrt{\frac{4M^2}{m+n_k}\log\frac{2|\mathcal{H}|}{\delta}}+\frac{M^2}{m+n_k}. \qquad \text{(D.6)}$$

Corollary D.17. *There is an absolute constant γ_1, for any $k=0,\dots,K$, any $\delta > 0$, if $\frac{2M\log\frac{|\mathcal{H}|}{\delta}}{m+n_k} \geq \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} \geq \frac{M}{2})$, then with probability at least $1-\delta$, for all $h_1,h_2 \in \mathcal{H}$, the following statements* *hold:*

$$\big|\big(l(h_1;S_k,M)-l(h_2;S_k,M)\big)-\big(l(h_1)-l(h_2)\big)\big| \le \gamma_1\frac{M}{m+n_k}\log\frac{|\mathcal{H}|}{\delta}+\gamma_1\frac{M^2}{(m+n_k)^{\frac{3}{2}}}\sqrt{\log\frac{|\mathcal{H}|}{\delta}} \tag{D.7}$$

$$+\gamma_1\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{m+n_k}\hat{V}(h_1,h_2;S_k,M)};$$

$$l(h_1;S_k,M)\le 2l(h_1)+\gamma_1\frac{M}{m+n_k}\log\frac{|\mathcal{H}|}{\delta}. \tag{D.8}$$

Proof. Let event E be the event that (D.1), (D.2), and (D.5) hold for all $h_1,h_2\in\mathcal{H}$ with confidence $1-\frac{\delta}{3}$ respectively. Assume E happens (whose probability is at least $1-\delta$).

(D.7) is immediate from (D.1) and (D.5).

For the proof of (D.8), apply (D.2) to h_1, we get

$$l(h_1;S_k,M)\le l(h_1)+\frac{10\log\frac{6|\mathcal{H}|}{\delta}}{3(m+n_k)}M+\sqrt{\frac{4\log\frac{6|\mathcal{H}|}{\delta}}{m+n_k}\mathrm{V}(h_1;k,M)}.$$

Now, $\mathrm{V}(h_1;k,M)=\mathbb{E}\left[w_k(X)\mathbb{1}\{h_1(X)\neq Y\}\mathbb{1}\{w_k(X)\le M\}\right]\le M\mathbb{E}[\mathbb{1}\{h_1(X)\neq Y\}]$, so $\sqrt{\frac{4\log\frac{6|\mathcal{H}|}{\delta}}{m+n_k}\mathrm{V}(h_1;k,M)}\le\sqrt{\frac{4M\log\frac{6|\mathcal{H}|}{\delta}}{m+n_k}l(h_1)}\le l(h_1)+\frac{M\log\frac{6|\mathcal{H}|}{\delta}}{(m+n_k)}$ where the last inequality follows from $\sqrt{ab}\le\frac{a+b}{2}$ for $a,b\ge 0$, and (D.8) thus follows. □

D.3 Technical Lemmas

For any $0 \leq k < K$ and $\delta > 0$, define event $\mathcal{E}_{k,\delta}$ to be the event that the conclusions of Theorem D.14 and Corollary D.16 hold for k with confidence $1-\delta/2$ respectively. We have $\mathbb{P}(\mathcal{E}_{k,\delta}) \geq 1-\delta$, and that $\mathcal{E}_{k,\delta}$ implies inequalities (D.7) and (D.8).

Recall that $\sigma_1(k,\delta,M) = \frac{M}{m+n_k}\log\frac{|\mathcal{H}|}{\delta} + \frac{M^2}{(m+n_k)^{\frac{3}{2}}}\sqrt{\log\frac{|\mathcal{H}|}{\delta}}; \sigma_2(k,\delta) = \frac{1}{m+n_k}\log\frac{|\mathcal{H}|}{\delta}; \delta_k = \frac{\delta}{2(k+1)(k+2)}$.

We first present a lemma which can be used to guarantee that $h^\star$ stays in candidate sets with high probability by induction.

Lemma D.18. *For any $k = 0, \ldots K$, any $\delta > 0$, any $M \geq 1$ such that $\frac{2M\log\frac{|\mathcal{H}|}{\delta}}{m+n_k} \geq \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} \geq \frac{M}{2})$, on event $\mathcal{E}_{k,\delta}$, if $h^\star \in C_k$, then,*

$$l(h^\star;\tilde{S}_k,M) \leq l(\hat{h}_k;\tilde{S}_k,M) + \gamma_1\sigma_1(k,\delta,M) + \gamma_1\sqrt{\sigma_2(k,\delta)\hat{V}(h^\star,\hat{h}_k;\tilde{S}_k,M)}.$$

Proof.

$$\begin{aligned}
&l(h^\star;\tilde{S}_k,M) - l(\hat{h}_k;\tilde{S}_k,M)\\
=&l(h^\star;S_k,M) - l(\hat{h}_k;S_k,M)\\
\leq&\gamma_1\sigma_1(k,\delta,M) + \gamma_1\sqrt{\sigma_2(k,\delta)\hat{V}(h^\star,\hat{h}_k;S_k,M)}\\
=&\gamma_1\sigma_1(k,\delta,M) + \gamma_1\sqrt{\sigma_2(k,\delta)\hat{V}(h^\star,\hat{h}_k;\tilde{S}_k,M)}
\end{aligned}$$

The first and the second equalities follow by Lemma D.2. The inequality follows by Corollary D.17. □

Next, we present a lemma to bound the probability mass of the disagreement region of candidate sets.

Lemma D.19. *Let $\hat{h}_{k,M} = \arg\min_{h \in C_k} l(h;\tilde{S}_k,M)$, and $C_{k+1}(\delta,M) := \{h \in C_k \mid l(h;\tilde{S}_k,M) \leq l(\hat{h}_{k,M};\tilde{S}_k,M) + \gamma_1\sigma_1(k,\delta,M) + \gamma_1\sqrt{\sigma_2(k,\delta)\hat{V}(h,\hat{h}_{k,M};\tilde{S}_k,M)}\}$. There is an absolute constant $\gamma_2 > 1$ such that for any $k = 0,\ldots,K$, any $\delta > 0$, any $M \geq 1$ such that $\frac{2M\log\frac{|\mathcal{H}|}{\delta}}{m+n_k} \geq \mathbb{P}(\frac{m+n_k}{mQ_0(X)+n_k} \geq \frac{M}{2})$, on event $\mathcal{E}_{k,\delta}$, if $h^\star \in C_k$, then for all $h \in C_{k+1}(\delta,M)$,*

$$l(h) - l(h^\star) \leq \gamma_2\sigma_1(k,\delta,M) + \gamma_2\sqrt{\sigma_2(k,\delta)Ml(h^\star)}.$$

Proof. For any $h \in C_{k+1}(\delta,M)$, we have

$$\begin{aligned}
&l(h) - l(h^\star) \\
\leq& l(h;S_k,M) - l(h^\star;S_k,M) + \frac{10M\log\frac{4|\mathcal{H}|}{\delta}}{3(m+n_k)} + \sqrt{4\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}} \\
=& l(h;\tilde{S}_k,M) - l(h^\star;\tilde{S}_k,M) + \frac{10M\log\frac{4|\mathcal{H}|}{\delta}}{3(m+n_k)} + \sqrt{4\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}} \\
=& l(h;\tilde{S}_k,M) - l(\hat{h}_{k,M};\tilde{S}_k,M) + l(\hat{h}_{k,M};\tilde{S}_k,M) - l(h^\star;\tilde{S}_k,M) \\
&+ \frac{10M\log\frac{4\mathcal{H}|}{\delta}}{3(m+n_k)} + \sqrt{4\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}} \\
\leq& \gamma_1\sigma_1(k,\delta,M) + \gamma_1\sqrt{\sigma_2(k,\delta)\hat{\mathrm{V}}(h,\hat{h}_{k,M};\tilde{S}_k,M)} + \frac{10M\log\frac{4|\mathcal{H}|}{\delta}}{3(m+n_k)} + \sqrt{4\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}
\end{aligned} \tag{D.9}$$

where the first equality follows from Lemma D.2, the first inequality follows from Theorem D.14, and the second inequality follows from the definition of $C_k(\delta,M)$ and that $l(\hat{h}_{k,M};\tilde{S}_k,M) \leq l(h^\star;\tilde{S}_k,M)$.

Next, we upper bound $\sqrt{\hat{\mathrm{V}}(h,\hat{h}_{k,M};\tilde{S}_k,M)}$. We have

$$\begin{aligned}\sqrt{\hat{\mathrm{V}}(h,\hat{h}_{k,M};\tilde{S}_k,M)} &\leq \sqrt{\hat{\mathrm{V}}(h,h^\star;\tilde{S}_k,M)+\hat{\mathrm{V}}(h^\star,\hat{h}_{k,M};\tilde{S}_k,M)}\\ &\leq \sqrt{\hat{\mathrm{V}}(h,h^\star;\tilde{S}_k,M)}+\sqrt{\hat{\mathrm{V}}(h^\star,\hat{h}_{k,M};\tilde{S}_k,M)}\end{aligned}$$

where the first inequality follows from the triangle inequality $\hat{\mathrm{V}}(h,\hat{h}_{k,M};\tilde{S}_k,M)\leq\hat{\mathrm{V}}(h,h^\star;\tilde{S}_k,M)+\hat{\mathrm{V}}(h^\star,\hat{h}_{k,M};\tilde{S}_k,M)$ and the second follows from the fact that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for $a,b\geq 0$.

For the first term, we have $\sqrt{\hat{\mathrm{V}}(h,h^\star;\tilde{S}_k,M)}=\sqrt{\hat{\mathrm{V}}(h,h^\star;S_k,M)}\leq\sqrt{\mathrm{V}(h,h^\star;k,M)}+\sqrt{\frac{4M^2}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}+\frac{M^2}{m+n_k}$ by Corollary D.16.

For the second term, we have

$$\begin{aligned}\sqrt{\hat{\mathrm{V}}(h^\star,\hat{h}_{k,M};\tilde{S},M)} &\leq \sqrt{M(l(h^\star;\tilde{S}_k,M)+l(\hat{h}_{k,M};\tilde{S}_k,M))}\\ &\leq \sqrt{2Ml(h^\star;\tilde{S}_k,M)}\\ &\leq \sqrt{2Ml(h^\star;S_k,M)}\\ &\leq \sqrt{2M(2l(h^\star)+\gamma_1\frac{M}{m+n_k}\log\frac{|\mathcal{H}|}{\delta})}\\ &\leq \sqrt{\frac{2\gamma_1 M^2}{m+n_k}\log\frac{|\mathcal{H}|}{\delta}}+2\sqrt{Ml(h^\star)}\end{aligned}$$

where the first inequality follows from the inequality $w_k^2(X)Z\mathbb{1}\{h^\star(X)\neq\hat{h}_{k,M}(X)\}\mathbb{1}[w_k(X)\leq M]$ $\leq M(w_k(X)Z\mathbb{1}\{h^\star(X)\neq Y\}+w_k(X)Z\mathbb{1}\{\hat{h}_{k,M}(X)\neq Y\})$, the second inequality follows since $l(\hat{h}_{k,M};\tilde{S}_k,M)\leq l(h^\star;\tilde{S}_k,M)$, the third follows by Lemma D.3 since we assume $h^\star\in C_k$, the fourth follows by Corollary D.17, and the last follows by $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$.

Therefore, $\sqrt{\hat{\mathrm{V}}(h,\hat{h}_{k,M};\tilde{S}_k,M)}\leq\sqrt{\mathrm{V}(h,h^\star;k,M)}+(2+\sqrt{2\gamma_1})\sqrt{\frac{M^2}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}+\frac{M^2}{m+n_k}+$

$2\sqrt{Ml(h^\star)}$. Continuing (D.9), we have

$$l(h)-l(h^\star) \le (\frac{10}{3}+3\gamma_1+2\sqrt{2}\gamma_1^{\frac{3}{2}})\frac{M}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}+\gamma_1\frac{M^2}{(m+n_k)^{\frac{3}{2}}}\sqrt{\log\frac{4|\mathcal{H}|}{\delta}}$$
$$+(\gamma_1+2)\sqrt{\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}+2\gamma_1\sqrt{\frac{Ml(h^\star)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}.$$

Now, because $w_k^2(X)Z\mathbb{1}\{h^\star(X)\neq\hat{h}_k(X)\}\mathbb{1}[w_k(X)\le M]\le Mw_k(X)Z\mathbb{1}\{h^\star(X)\neq Y\}+Mw_k(X)Z\mathbb{1}\{\hat{h}_k(X)\neq Y\}$, we have that $\sqrt{\frac{\mathrm{V}(h^\star,h;k,M)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}\le\sqrt{\frac{M(l(h)-l(h^\star)+2l(h^\star))}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}\le\sqrt{\frac{M(l(h)-l(h^\star))}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}+\sqrt{\frac{2Ml(h^\star)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}$ where the second follows by $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$.

Thus, $l(h)-l(h^\star)\le(\frac{10}{3}+3\gamma_1+2\sqrt{2}\gamma_1^{\frac{3}{2}})\frac{M}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}+\gamma_1\frac{M^2}{(m+n_k)^{\frac{3}{2}}}\sqrt{\log\frac{4|\mathcal{H}|}{\delta}}+(2\gamma_1+\sqrt{2}\gamma_1+2\sqrt{2})\sqrt{\frac{Ml(h^\star)}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}+(\gamma_1+2)\sqrt{\frac{M(l(h)-l(h^\star))}{m+n_k}\log\frac{4|\mathcal{H}|}{\delta}}$.

The result follows by applying Lemma D.1 to $l(h)-l(h^\star)$. □

D.4 Proofs for Section 7.3.2

Proof. (of Theorem 7.9) Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K}\mathcal{E}_{k,\delta_k}$. By a union bound, $\mathbb{P}(\mathcal{E}^{(0)})\ge 1-\delta/2$. On event $\mathcal{E}^{(0)}$, by induction and Lemma D.18, for all $k=0,\dots,K$, $h^\star\in C_k$.

$$
\begin{aligned}
l(\hat{h}) - l(h^\star) \leq & l(\hat{h}; S_K, M_K) - l(h^\star; S_K, M_K) + \gamma_1 \sigma_1(K, \delta_K, M_K) + \gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(\hat{h}, h^\star; S_K, M_K)} \\
= & l(\hat{h}; \tilde{S}_K, M_K) - l(h^\star; \tilde{S}_K, M_K) + \gamma_1 \sigma_1(K, \delta_K, M_K) + \gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(\hat{h}, h^\star; \tilde{S}_K, M_K)} \\
\leq & l(\hat{h}; \tilde{S}_K, M_K) + \gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(\hat{h}; \tilde{S}_K, M_K)} \\
& - l(h^\star; \tilde{S}_K, M_K) - \gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(h^\star; \tilde{S}_K, M_K)} \\
& + \gamma_1 \sigma_1(K, \delta_K, M_K) + 2\gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(h^\star; \tilde{S}_K, M_K)} \\
\leq & \gamma_1 \sigma_1(K, \delta_K, M_K) + 2\gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(h^\star; \tilde{S}_K, M_K)} \\
\leq & \gamma_1 \sigma_1(K, \delta_K, M_K) + 2\gamma_1 \sqrt{\sigma_2(K, \delta_K) \hat{\mathrm{V}}(h^\star; S_K, M_K)} \\
\leq & 3\gamma_1 \sigma_1(K, \delta_K, M_K) + 2\gamma_1 \sqrt{\sigma_2(K, \delta_K) \mathrm{V}(h^\star; K, M_K)}
\end{aligned}
$$

where the equality follows from Lemma D.2, the first inequality follows from Corollary D.17, the second follows since $\sqrt{\hat{\mathrm{V}}(\hat{h}, h^\star; \tilde{S}_K, M_K)} \leq \sqrt{\hat{\mathrm{V}}(\hat{h}; \tilde{S}_K, M_K) + \hat{\mathrm{V}}(h^\star; \tilde{S}_K, M_K)} \leq \sqrt{\hat{\mathrm{V}}(\hat{h}; \tilde{S}_K, M_K)} + \sqrt{\hat{\mathrm{V}}(h^\star; \tilde{S}_K, M_K)}$, the third follows from the definition of $\hat{h}$, the forth follows from Lemma D.3, and the last follows from Corollary D.16. □

Proof. (of Theorem 7.11) Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K} \mathcal{E}_{k,\delta_k}$. On this event, by induction and Lemma D.18, for all $k = 0, \dots, K-1$, $h^\star \in C_k$, and consequently by Lemma D.19, $D_{k+1} \subseteq \mathrm{DIS}(B(h^\star, 2\nu + \varepsilon_k))$ where $\varepsilon_k = \gamma_2 \sigma_1(k, \delta_k, M_k) + \gamma_2 \sqrt{\sigma_2(k, \delta_k) M_k \nu}$.

For any $k = 0, \dots K-1$, define the number of label queries at iteration k to be $U_k := \sum_{t=m+n_k+1}^{m+n_{k+1}} Z_t \mathbb{1}\{X_t \in D_{k+1}\}$ where the RHS is a sum of i.i.d. Bernoulli random variables with expectation $\mathbb{E}[Z_t \mathbb{1}\{X_t \in D_{k+1}\}] = \mathbb{P}(D_{k+1} \cap \{x : Q_0(x) < \frac{2n_{k+1}}{m}\})$ since $Z_t = Q_{k+1}(x) = \mathbb{1}\{2n_{k+1} - mQ_0(x) > 0\}$ by Proposition D.7. A Bernstein inequality implies that on an event $\mathcal{E}^{(1,k)}$ of probability at least $1 - \delta_k/2$, $U_k \leq 2\tau_{k+1} \mathbb{P}(D_{k+1} \cap \{x : Q_0(x) < \frac{2n_{k+1}}{m}\}) + 2\log\frac{4}{\delta_k}$.

Define $\mathcal{E}^{(1)} := \bigcap_{k=0}^{K-1} \mathcal{E}^{(1,k)}$, and $\mathcal{E}^{(2)} := \mathcal{E}^{(0)} \cap \mathcal{E}^{(1)}$. By a union bound, we have

$\mathbb{P}(\mathcal{E}^{(2)}) \geq 1-\delta$. Now, on event $\mathcal{E}^{(2)}$, for any $k < K$, $D_{k+1} \subseteq \text{DIS}(B(h^\star, 2\nu+\varepsilon_k))$, so by Lemma D.5 $\mathbb{P}(D_{k+1} \cap \{x : Q_0(x) < \frac{2n_{k+1}}{m}\}) \leq (2\nu+\varepsilon_k)\tilde{\theta}(2\nu+\varepsilon_k, \frac{2n_{k+1}}{m})$. Therefore, the total number of label queries

$$\begin{aligned}
\sum_{k=0}^{K-1} U_k \leq & \tau_1 + \sum_{k=1}^{K-1} 2\tau_{k+1} \mathbb{P}(D_{k+1} \cap \{x : Q_0(x) < \frac{2n_{k+1}}{m}\}) + 2K \log \frac{4}{\delta_K} \\
\leq & 1 + 2\sum_{k=1}^{K-1} \tau_{k+1}(2\nu+\varepsilon_k)\tilde{\theta}(2\nu+\varepsilon_k, \frac{2n_{k+1}}{m}) + 2K \log \frac{4}{\delta_K} \\
\leq & 1 + 2K\log\frac{4}{\delta_K} + 2\tilde{\theta}(2\nu+\varepsilon_{K-1}, \frac{2n}{m}) \cdot \Bigg(2n\nu \\
& + \gamma_2 \sum_{k=1}^{K-1} (\frac{\tau_{k+1} M_k}{m+n_k} \log\frac{|\mathcal{H}|}{\delta_k} + \frac{\tau_{k+1} M_k^2}{(m+n_k)^{\frac{3}{2}}} \sqrt{\log\frac{|\mathcal{H}|}{\delta_k}} + \tau_{k+1}\sqrt{\frac{M_k}{m+n_k} \nu \log\frac{|\mathcal{H}|}{\delta_k}}) \Bigg).
\end{aligned}$$

Recall that $\alpha = \frac{m}{n}, \tau_k = 2^k$, $\xi = \min_{1\leq k \leq K}\{M_k / \frac{m+n_k}{mq_0+n_k}\}$, $\bar{M} = \max_{1\leq k\leq K} M_k$. We have $\sum_{k=1}^{K-1} \frac{\tau_{k+1} M_k}{m+n_k} \leq \sum_{k=1}^{K-1} \frac{\xi\tau_k}{mq_0+n_k} \leq \sum_{k=1}^{K} \frac{\xi n_k}{\alpha n_k q_0 + n_k} \leq \frac{K\xi}{\alpha q_0+1}$ where the first inequality follows as $\frac{M_k}{m+n_k} \leq \frac{\xi}{mq_0+n_k}$, and the second follows by $m = n\alpha \geq n_k\alpha$. Besides, $\sum_{k=1}^{K-1} \frac{\tau_k M_k^2}{(m+n_k)^{\frac{3}{2}}} \leq \sum_{k=1}^{K-1} \frac{\tau_k M_k \xi}{\sqrt{m+n_k}(mq_0+n_k)}$ $\leq \sum_{k=1}^{K-1} \frac{\bar{M}\xi}{\sqrt{m+n_k}} \leq \frac{K\bar{M}\xi}{\sqrt{n\alpha}}$ where the first inequality follows as $\frac{M_k}{m+n_k} \leq \frac{\xi}{mq_0+n_k}$, and the second follows as $M_k \leq \bar{M}$ and $\tau_k \leq mq_0+n_k$. Finally, $\sum_{k=1}^{K} \tau_k \sqrt{\frac{M_k}{m+n_k}} \leq \sum_{k=1}^{K} \sqrt{\frac{\tau_k \xi}{\alpha q_0+1}} \leq \sqrt{\frac{n\xi}{\alpha q_0+1}}$ where the first inequality follows as $\frac{M_k}{m+n_k} \leq \frac{\xi}{mq_0+n_k}$ and $mq_0+n_k \geq \tau_k(\alpha q_0+1)$.

Therefore,

$$\sum_{k=0}^{K-1} U_k \leq 1 + 2K \log \frac{4}{\delta_K} + 2\tilde{\theta}(2\nu + \varepsilon_{K-1}, \frac{2n}{m}) \Bigg(2n\nu$$
$$+ \gamma_2 (\frac{K\xi}{\alpha q_0 + 1} \log \frac{K^2|\mathcal{H}|}{\delta} + \frac{K\bar{M}\xi}{\sqrt{n\alpha}} \sqrt{\log \frac{K^2|\mathcal{H}|}{\delta}} + \sqrt{\frac{n\xi\nu}{\alpha q_0 + 1} \log \frac{K^2|\mathcal{H}|}{\delta}}) \Bigg) .$$

□

D.5 Proofs for Sections 7.2

Theorem 7.1 and Corollary 7.6 are immediate from the following theorem.

Theorem D.20. *Let* $\hat{h}_M = \arg\min_{h \in \mathcal{H}} l(h; S, M) + \sqrt{\frac{\lambda}{m} \hat{V}(h; S, M)}$. *For any* $\delta > 0$, $M \geq 1$, $\lambda \geq 4 \log \frac{|\mathcal{H}|}{\delta}$, *with probability at least* $1 - \delta$ *over the choice of S,*

$$l(\hat{h}_M) - l(h^\star) \leq \frac{2\lambda M}{m} + \frac{16M}{3m} \log \frac{|\mathcal{H}|}{\delta} + \frac{M^2}{m^{\frac{3}{2}}} \sqrt{4 \log \frac{|\mathcal{H}|}{\delta}}$$
$$+ \sqrt{\frac{\lambda}{m} \mathbb{E} \frac{\mathbb{1}\{h^\star(X) \neq Y\}}{Q_0(X)} \mathbb{1}[\frac{1}{Q_0(X)} \leq M]} + \mathbb{P}_X(\frac{1}{Q_0(X)} > M). \quad \text{(D.10)}$$

Proof. The proof is similar to the proofs for Theorem 7.9 and D.14, and is omitted. □

Bibliography

[AB09] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[ABHU15] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Proceedings*, pages 167–190. JMLR.org, 2015.

[ABHZ16] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2016*, 2016.

[ABL14] P. Awasthi, M-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *STOC*, 2014.

[ABSJ17] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. Effective evaluation using logged bandit feedback from multiple loggers. *arXiv preprint arXiv:1703.06180*, 2017.

[ABSS93] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.

[AHK$^+$14] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

[AI16] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[AL88] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988.

[Ang88] Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

[AZvdS19] Onur Atan, William R. Zame, and Mihaela van der Schaar. Sequential patient recruitment and allocation for adaptive clinical trials. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1891–1900. PMLR, 16–18 Apr 2019.

[BBL06a] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *ICML*, 2006.

[BBL06b] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

[BBL09] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.

[BBZ07] M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.

[BDBC$^+$10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[BDL09] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

[BFKV98] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[BH12] Maria-Florina Balcan and Steve Hanneke. Robust interactive learning. In *COLT*, 2012.

[BHLZ10] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

[BHLZ16] Alina Beygelzimer, Daniel J Hsu, John Langford, and Chicheng Zhang. Search improves label for active learning. In *Advances in Neural Information Processing Systems*, pages 3342–3350, 2016.

[BL13] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[BPQC$^+$13] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[BS79] P Borjesson and C-E Sundberg. Simple approximations of the error function q (x) for communications applications. *IEEE Transactions on Communications*, 27(3):639–643, 1979.

[CAL94] D. A. Cohn, L. E. Atlas, and R. E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

[CHK17] Lin Chen, Hamed Hassani, and Amin Karbasi. Near-optimal active learning of halfspaces via query synthesis in the noisy setting. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[CL11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[CMM10] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[CMMR12] Jean Cornuet, JEAN-MICHEL MARIN, Antonietta Mira, and Christian P Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

[CN07] Rui Castro and Robert D. Nowak. Minimax bounds for active learning. In *COLT*, pages 5–19, 2007.

[CN08] Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[CST00] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. 2000.

[Dan15] Amit Daniely. Complexity theoretic limitations on learning halfspaces. *arXiv preprint arXiv:1505.05800*, 2015.

[Das05] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[Das11] S. Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19), 2011.

[DHM07] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

[DKM05] Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 249–263, 2005.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

[DV04] John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 315–320. ACM, 2004.

[ERR16] Mehdi Elahi, Francesco Ricci, and Neil Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50, 2016.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 563–574. IEEE, 2006.

[FSST97] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[FZ12] Meng Fang and Xingquan Zhu. I don't know the label: Active learning with blind knowledge. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2238–2241. IEEE, 2012.

[GCB09] Andrew Guillory, Erick Chastain, and Jeff Bilmes. Active learning as non-convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 201–208, 2009.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[HAH$^+$15] 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.

[Han07] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[Han09] S. Hanneke. Adaptive rates of convergence in active learning. In *COLT*, 2009.

[Han14] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[Heg95] Tibor Hegedűs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 108–117. ACM, 1995.

[HLR16] Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval*, 10(1):1–117, 2016.

[Hsu10] D. Hsu. *Algorithms for Active Learning*. PhD thesis, UC San Diego, 2010.

[HY15] Steve Hanneke and Liu Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015.

[JL16] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 652–661. JMLR. org, 2016.

[JSS16] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

[KAH$^+$17] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé, III, and John Langford. Active learning for cost-sensitive classification. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1915–1924, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Kal17] Nathan Kallus. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pages 1789–1798, 2017.

[KFR$^+$15] Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4343–4352. IEEE, 2015.

[KGR$^+$15] David Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, and Yan Liu. Hierarchical active transfer learning. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 514–522. SIAM, 2015.

[KK14] Adam Klivans and Pravesh Kothari. Embedding Hard Learning Problems Into Gaussian Space. In Klaus Jansen, José D. P. Rolim, Nikhil R. Devanur, and Cristopher Moore, editors, *Approximation, Randomization, and Combinatorial*

Optimization. Algorithms and Techniques (APPROX/RANDOM 2014), volume 28 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 793–809, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[KL11] Nikos Karampatziakis and John Langford. Online importance weight aware updates. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 392–399. AUAI Press, 2011.

[KLS09] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

[KM18] Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886, 2018.

[KMT93] Sanjeev R Kulkarni, Sanjoy K Mitter, and John N Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993.

[Kol10] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *JMLR*, 2010.

[KRG18] Saikishore Kalloori, Francesco Ricci, and Rosella Gennari. Eliciting pairwise preferences in recommender systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 329–337. ACM, 2018.

[KSH$^+$16] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016.

[LCKG15] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934. ACM, 2015.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[LG94] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[Lon95] Philip M Long. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[LWD$^+$19] C. Li, X. Wang, W. Dong, J. Yan, Q. Liu, and H. Zha. Joint active learning with feature selection via cur matrix decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1382–1396, June 2019.

[LY13] Yuan-Chuan Li and Cheh-Chih Yeh. Some equivalent forms of bernoulli's inequality: A survey. *Applied Mathematics*, 4(07):1070, 2013.

[Min12] Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(Jan):67–90, 2012.

[MLBP16] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. Offline evaluation of online reinforcement learning algorithms. 2016.

[MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, pages 2326–2366, 2006.

[Mon06] Claire Monteleoni. Efficient algorithms for general active learning. In *International Conference on Computational Learning Theory*, pages 650–652. Springer, 2006.

[MP09] A Maurer and M Pontil. Empirical bernstein bounds and sample variance penalization. In *COLT 2009-The 22nd Conference on Learning Theory*, 2009.

[MS54] TS Motzkin and IJ Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6(3):393–404, 1954.

[MTSVDH15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[ND17] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

[NJC15] Mohammad Naghshvar, Tara Javidi, and Kamalika Chaudhuri. Bayesian active learning with non-persistent noise. *IEEE Transactions on Information Theory*, 61(7):4080–4098, 2015.

[Now11] R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

[NS04] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79. ACM, 2004.

[OZ00] Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

[RB16] Aaditya Ramdas and Akshay Balsubramani. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

[RF17] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RR11a] Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

[RR11b] Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

[SCR08] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[Set10] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

[SH16] Sivan Sabato and Tom Hess. Interactive algorithms: from pool to stream. In *Conference on Learning Theory*, pages 1419–1439, 2016.

[SJ15a] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[SJ15b] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, pages 3231–3239, 2015.

[SJS17] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[SKM07] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[SLLK10] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.

[SOS92] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[SRD$^+$11] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[SS18] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[SSS$^+$16] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352*, 2016.

[SSS$^+$19] Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. Active learning for decision-making from imbalanced observational data. *arXiv preprint arXiv:1904.05268*, 2019.

[SYL$^+$18] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, 2018.

[TB16] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[TD17] Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *ICML*, pages 3444–3452, 2017.

[Tho15] Philip S Thomas. Safe reinforcement learning. 2015.

[TK01] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[Tsy04] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

[Tsy08] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[TTG15] Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, pages 3000–3006, 2015.

[Val84] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[Vap98] Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[VC71] VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

[VG95] Eric Veach and Leonidas J Guibas. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 419–428. ACM, 1995.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[vw] Vowpal Wabbit. https://github.com/JohnLangford/vowpal_wabbit/.

[WAD17] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597, 2017.

[WS16] Yining Wang and Aarti Singh. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *AAAI*, 2016.

[XZM$^+$17] Yichong Xu, Hongyang Zhang, Kyle Miller, Aarti Singh, and Artur Dubrawski. Noise-tolerant interactive learning using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2431–2440, 2017.

[YCJ15] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from noisy and abstention feedback. In *Communication, Control, and Computing (Allerton), 2015 53th Annual Allerton Conference on*. IEEE, 2015.

[YCJ16] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. In *Advances in Neural Information Processing Systems*, pages 2128–2136, 2016.

[YCJ18] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning with logged data. In *International Conference on Machine Learning*, pages 5517–5526, 2018.

[YCJ19] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. The label complexity of active learning from observational data. *arXiv preprint arXiv:1905.12791*, 2019.

[YZ17] Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems*, pages 1056–1066, 2017.

[Zad04] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.

[ZAI$^+$19] Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, and Sahand Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7335–7344, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[ZC14] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.

[ZC15] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems*, pages 703–711, 2015.

[ZJL$^+$16] Zihan Zhang, Xiaoming Jin, Lianghao Li, Guiguang Ding, and Qiang Yang. Multi-domain active learning for recommendation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[ZS13] Andre M Zubkov and Aleksandr A Serov. A complete proof of universal inequalities for the distribution function of the binomial law. *Theory of Probability & Its Applications*, 57(3):539–544, 2013.