DISTRIBUTION-SPECIFIC AGNOSTIC CONDITIONAL CLASSIFICATION WITH HALFSPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study "selective" or "conditional" classification problems under an agnostic setting. Classification tasks commonly focus on modeling the relationship between features and categories that captures the vast majority of data. In contrast to common machine learning frameworks, conditional classification intends to model such relationships only on a subset of the data defined by some selection rule. Most work on conditional classification either solves the problem in a realizable setting or does not guarantee the error is bounded compared to an optimal solution. In this work, we consider selective/conditional classification by sparse linear classifiers for subsets defined by halfspaces, and give both positive as well as negative results for Gaussian feature distributions. On the positive side, we present the first PAC-learning algorithm for homogeneous halfspace selectors with error guarantee $\tilde{O}(\sqrt{\mathrm{opt}})$, where opt is the smallest conditional classification error over the given class of classifiers and homogeneous halfspaces. On the negative side, we find that, under cryptographic assumptions, approximating the conditional classification loss within a small additive error is computationally hard even under Gaussian distribution. We prove that approximating conditional classification is at least as hard as approximating agnostic classification in both additive and multiplicative form.

1 INTRODUCTION

Classification is the task of modeling the relationship between some features and membership in some category. Classification tasks are common across various fields, such as spam detection (classifying emails as "spam" or "not spam"), image recognition (identifying objects like "cat" or "dog"), and medical diagnosis (predicting whether a patient has a certain condition or not). Standard classification approaches seek to model the whole data distribution. By contrast, we consider the problems where a **better classifier** exists on a subset of the data. In particular, we will consider cases in which classifiers are sparse linear functions (or more generally, any small set of functions), and subsets are described by **selector** functions, given here by homogeneous halfspaces.

We study the distribution-specific PAC-learnability (Kearns et al., 1994) of the class of classifier-selector pairs in the presence of adversarial label noise. In the literature, this problem is known as "conditional" classification, but it is also part of a family of problems that are generally known as "selective" classification.

1.1 BACKGROUND AND MOTIVATION

The first "selective classification" problem was introduced decades ago (Chow, 1957; 1970). The focus was on finding Bayes classifiers for the case where the data distribution is fully known. The appeal of effective selective classification is clear in applications where partial domain coverage is acceptable, or in scenarios where achieving extremely low risk is essential but unattainable with standard classification methods. Classification tasks in medical diagnosis and bioinformatics often fall into this category (Khan et al., 2001; Hanczar & Dougherty, 2008).

El-Yaniv et al. (2010) gave a thorough theoretical analysis for selective classification based on a "risk-coverage" model. They proved that, for the optimal classifier and selector, there exists a natural trade-off between the performance of the classifier on the selected subset and the size of the subset.

Prior work has either considered the "realizable" case (El-Yaniv & Wiener, 2012; Gangrade et al., 2021), where there exists a classifier-selector pair that does not make any errors, or endowed the learner with a rejection mechanism using heuristic rules or confidence scores (Geifman & El-Yaniv, 2017; Pugnana & Ruggieri, 2023). For the "agnostic" case, where no perfect classifier-selector pair exists, few works had been done on model-based selective learning (Wiener & El-Yaniv, 2011; 2015; Gelbhart & El-Yaniv, 2019). More importantly, these works do not guarantee both computational efficiency together with good performance with respect to the optimal classifier and selector.

We consider a more general formulation of **agnostic selective classification** under the PAC-learning semantics in Definition 1.1. In particular, we do not make any assumptions on the labels while the performance of the learned classifier and selector are guaranteed to be close to the optimal solution.

Definition 1.1 (Agnostic Conditional Classification)**.** *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$, $\mathcal{C}$ be a finite class of classifiers on $\mathbb{R}^d \times \{0,1\}$, and $\mathcal{H} = \{S \subseteq \mathbb{R}^d \mid \Pr_{\mathcal{D}}\{S\} \in [a,b]\}$ for $0 \leq a \leq b \leq 1$. Suppose $\min_{S\in\mathcal{H}, c\in\mathcal{C}} \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{y \neq c(\mathbf{x}) \mid \mathbf{x} \in S\} = \mathrm{opt}$, for some $C > 1$. A C-approximate learning algorithm (or an algorithm with approximation factor C), given sample access to $\mathcal{D}$, outputs an $S' \in \mathcal{H}$ such that $\min_{c\in\mathcal{C}} \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{y \neq c(\mathbf{x}) \mid \mathbf{x} \in S'\} \leq C \cdot \mathrm{opt}$ with high probability.*

The imposed "population" bounds on the subsets $S \in \mathcal{H}$ are critical. On the one hand, the lower bound, $\Pr\{S\} \geq a$ can both prevent trivial optimal solutions such as $S' = \varnothing$ and make the selected subsets statistically meaningful. On the other hand, if the selector chooses a majority of the data, the performance advantage of the optimal solution of selective classification could vanish compared with that of the regular classification model (El-Yaniv et al., 2010; Hainline et al., 2019).

Consider a halfspace h, i.e., a subset of $\mathbb{R}^d$ such that the membership in h is defined by some linear threshold function. In this work, we wish to solve the problem of agnostic conditional classification with halfspace selectors under standard normal distributions described as follows.

Problem 1.2 (Distribution-Specific Agnostic Conditional Classification With Halfspaces)**.** *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginal on $\mathbb{R}^d$, $\mathcal{C}$ be a finite class of classifiers on $\mathbb{R}^d \times \{0,1\}$, and $\mathcal{H}$ be the class of halfspaces on $\mathbb{R}^d$ with population size in the range of $[a,b]$ for $0 \leq a \leq b \leq 1$. Suppose $\min_{h\in\mathcal{H}, c\in\mathcal{C}} \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{y \neq c(\mathbf{x}) \mid \mathbf{x} \in h\} = \mathrm{opt}$, how close to opt can a polynomial-time learning algorithm achieve on $\mathcal{H}$ with high probability?*

An algorithm for Problem 1.2 may be leveraged to perform conditional classification for large or infinite classes $\mathcal{C}$ by using an algorithm for list learning of classifiers for some richer class (Charikar et al., 2017), taking $\mathcal{C}$ in Problem 1.2 to be the list of classifiers produced by the list learning algorithm:

Definition 1.3 (Robust list learning)**.** *Let $\mathcal{D} = \alpha\mathcal{D}^* + (1-\alpha)\tilde{\mathcal{D}}$ for an* inlier *distribution $\mathcal{D}^*$ and* outlier *distribution $\tilde{\mathcal{D}}$ each supported on $\mathbb{R}^d \times \{0,1\}$, with $\alpha \in (0,1)$. A* robust list learning *algorithm for a class of Boolean classifiers $\mathcal{C}$, given α and parameters $\epsilon, \delta \in (0,1)$, and sample access to $\mathcal{D}$ such that for $(\mathbf{x}, b)$ in the support of $\mathcal{D}^*$, $b = c^*(\mathbf{x})$ for some $c^* \in \mathcal{C}$, runs in time $\mathrm{poly}\left(d, \frac{1}{\alpha}, \frac{1}{\epsilon}, \log\frac{1}{\delta}\right)$, and with probability $1-\delta$ returns a list of $\ell = \mathrm{poly}\left(d, \frac{1}{\alpha}, \frac{1}{\epsilon}, \log\frac{1}{\delta}\right)$ classifiers $\{h_1, \ldots, h_\ell\}$ such that for some h_i in the list, $\Pr_{\mathcal{D}^*}[h_i(\mathbf{x}) = c^*(\mathbf{x})] \geq 1-\epsilon$.*

As we will review, it is known in particular that, for sparse linear classifiers (with $s = O(1)$ nonzero coefficients), list learning from a sample of size $m = O(\frac{1}{\alpha\epsilon}(s\log d + \log\frac{1}{\delta}))$ is possible in time and list size $O((md)^s)$ (Juba, 2017; Mossel & Sudan, 2016).

1.2 Challenges Of Distribution-Specific Conditional Classification

Problem 1.2 is similar to **agnostic linear classification**, where we seek to minimize the classification error over the vast majority of data. In particular, it was clear that agnostic classification can be reduced to (distribution-free) conditional learning (Juba, 2017). Agnostic linear classification has been extensively studied over decades, and it is known to be computationally hard in both distribution-free (Kearns et al., 1994) and distribution-specific settings (Diakonikolas et al., 2023).

Despite the intractability of agnostic learning, numerous distribution-specific approximation schemes have been developed with approximation factor of $O(1/\sqrt{\mathrm{opt}})$ or even constants (Frei et al., 2021; Diakonikolas et al., 2020c; 2022; 2024; Shen, 2021). Given the similarity between agnostic linear classification and Problem 1.2, and that it was the only formal barrier known, it is natural to ask if we can leverage the existing techniques for standard agnostic classification in conditional classification.

However, it is not clear how these could lead to a meaningful error guarantee for conditional classification. Directly, Definition 1.1 (correspondingly, Problem 1.2) can be reduced to a "one-sided" classification problem, where we seek to minimize the error rate of the classifier on only one class. As the error rate could be extremely unbalanced across the classes, a constant factor approximation scheme for the agnostic linear classification problem may not yield approximation guarantees for the one-sided agnostic classification problem.

An analogous difficulty arose in "fairness auditing" (Kearns et al., 2018). In the problem of fairness auditing, instead of minimizing the classification error, we wish to verify some specific fairness criteria for a subset of the data. Kearns et al. (2018) showed that the auditing problem is equivalent to agnostic classification for any simple representation classes (including halfspaces) under distribution-free settings. Despite the similarity between these two problems, as well as the existence of constant factor approximation algorithms for agnostic linear classification under distributional assumptions, recent work by Hsu et al. (2024) showed there does not exist any nontrivial multiplicative factor approximation algorithm for auditing halfspace subgroup fairness even under Gaussian distributions. The connection in the distribution-free setting simply does not carry over to Gaussian data.

1.3 Our Contribution

Let opt be as defined in Problem 1.2 for $\mathcal{H}$ being the class of **homogeneous** halfspaces. Our first contribution is a polynomial-time $\tilde{O}(1/\sqrt{\mathrm{opt}})$-approximation algorithm to learn a pair of classifier and selector for Problem 1.2 with homogeneous halfspace selectors. This is the first polynomial-time algorithm for agnostic conditional/selective classification with a provable approximation guarantee w.r.t. the optimal solution.

Remark 1. *Even for homogeneous halfspace selectors, the imbalance of error rates between classes could still exist, as we will show in our hardness result that the difference between the error rates of different classes of the homogeneous halfspace always equals to the amount that the probability of either label deviates from* $1/2$*; see Lemma 4.4 for details.*

Our second contribution is a negative result for Problem 1.2. We show that agnostic conditional classification in Definition 1.1 is at least as hard as agnostic linear classification under any distribution. With the distribution-specific hardness result of agnostic linear classification (Diakonikolas et al., 2023), we prove that no polynomial-time algorithm can achieve an error guarantee of $\mathrm{opt} + O(1/\log^{1/2+\alpha} d)$ for any constant $\alpha > 0$ for Problem 1.2. We show more generally that approximating the conditional classification objective is at least as hard as approximating the regular classification objective.

Organization. In Section 2, we give some necessary background. We will present our algorithmic results in Section 3. The distribution-specific hardness result for conditional classification with general halfspaces is in Section 4. In the last section, we will discuss the limitations of our results and a few possible directions for extensions.

1.4 Related Works

Selective Learning. Besides the results we have mentioned above, there are many works on selective classification. For the realizable cases, El-Yaniv & Wiener (2012) reduced active learning to selective learning, and used this reduction to prove a exponential lower bound on label complexity for learning linear classifiers when using the *CAL algorithm*, which is one of the main strategies for active learning in the realizable setting. Gangrade et al. (2021) proposed a optimization-based selective learning framework that guarantees to maximize the classifiers' coverage with a specified one-side prediction error rate. They proved that any representation class with finite VC-dimension can be used successfully in their models. For the agnostic cases, Wiener & El-Yaniv (2011; 2015); Gelbhart & El-Yaniv (2019) presented a selective learning approach to learn a classifier-selector pair that is at least as competitive as the ERM of the non-selective learning task. However, the computation of both the classifier and selector in these methods relies on an agnostic learning oracle, and the selector function is not guaranteed to minimize the conditional classification error down to any approximation factor. Geifman & El-Yaniv (2017) proposed a method to design selector functions for any given deep neural network. Their selector is built upon a given heuristic scoring function for data examples and can provably guaranteed to achieve strong performance. Aside from the theoretical results, empirically, Pugnana & Ruggieri (2023) developed an model-agnostic learning algorithm to learn a

confidence-based selective classifier that seeks to minimize the AUC-based loss within the selected region and Geifman & El-Yaniv (2019) proposed the SelectiveNet architecture that simultaneously learns a pair of classifier and selector in a single neural networks with required coverage.

Conditional Learning. The problem of conditional learning (including conditional classification) incorporates two sub-problems, obtaining a finite list of classifiers as well as learning a classifier-selector pair out this finite list and some class of selector functions. For the former task, a series of positive results (Charikar et al., 2017; Kothari et al., 2018; Calderon et al., 2020; Bakshi & Kothari, 2021) have been obtained under the "list-decodable" setting of Definition 1.3. For the latter task, Juba (2016) introduced the problem of learning abduction, where they propose to learn a subset of the data distribution where e.g., no errors occur. In their work, they showed that subsets defined by k-DNFs can be efficiently learned in realizable cases without any distributional assumptions. Subsequent improvements were obtained for the agnostic setting (Zhang et al., 2017; Juba et al., 2018).

Learning To Abstain. Cortes et al. (2016) considered a different formulation of selective classification. Instead of optimizing the classification error conditioned the selected subgroup, they proposed to minimize the classification error jointly with the selector function while enforcing a cost for "abstaining". They designed a few convex surrogate losses to upper bound the joint classification loss in the setting that abstaining has a cost. Later works (Mao et al., 2024c;a;b) proposed new families of surrogate losses to approximate the classification loss with abstaining and proved various upper bounds classification error of any classifier-selector pair in terms of different surrogate loss measures for two different selective learning strategies.

2 Preliminaries

We use lowercase bold font characters to represent real vectors. In addition, subscripts will be used to index the coordinates of each vector $\mathbf{x} \in \mathbb{R}^d$, e.g., $\mathbf{x}_i$ represents the ith coordinate of vector $\mathbf{x}$. For $\mathbf{x} \in \mathbb{R}^d$, let $\|\mathbf{x}\|_p = (\sum_{i=1}^d \mathbf{x}_i^p)^{1/p}$ denote the l_p-norm of $\mathbf{x}$, and $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$ denote the normalized vector of $\mathbf{x}$. For any matrix $A \in \mathbb{R}^{m\times n}$, denote $\|A\|_{\mathrm{op}} = \max_{\|\mathbf{u}\|_2=1} \|A\mathbf{u}\|_2$. We will use $\langle \mathbf{x}, \mathbf{y}\rangle$ to represent the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x}^{\otimes k}$ to represent the outer product of $\mathbf{x} \in \mathbb{R}^d$ to the kth degree. Further, we will write $\mathbf{w}^\perp = \{\mathbf{u} \in \mathbb{R}^d \mid \langle \mathbf{u}, \mathbf{w}\rangle = 0\}$ as the orthogonal subspace of $\mathbf{w} \in \mathbb{R}^d$, and $\mathbf{x}_{\mathbf{w}^\perp} = (\mathrm{I} - \bar{\mathbf{w}}^{\otimes 2})\mathbf{x}$ as the projection of $\mathbf{x} \in \mathbb{R}^d$ onto $\mathbf{w}^\perp$. Additionally, we will use $\theta(\mathbf{u}, \mathbf{w})$ to denote the angle between two vector $\mathbf{u}, \mathbf{w} \in \mathbb{R}^d$.

We use $\mathcal{D}_{\mathbf{x}}$ to denote the marginal distribution of $\mathcal{D}$ on $\mathbf{x} \in \mathbb{R}^d$, $\Pr_{\mathcal{D}}\{E\}$ to denote the probability of an event E, and $\mathbb{E}_{\mathcal{D}}[X]$ to denote the expectation of some statistic X under distribution $\mathcal{D}$. In particular, for an empirical sample $\hat{\mathcal{D}} \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, we use $\mathbb{E}_{\hat{\mathcal{D}}}[X]$ to denote the empirical average of X, i.e., $\mathbb{E}_{X\sim\hat{\mathcal{D}}}[X] = 1/|\hat{\mathcal{D}}| \sum_{X\in\hat{\mathcal{D}}} X$. $\mathcal{N}^d(0,1)$ denotes the d-dimensional standard normal distribution. For simplicity, we may drop $\mathcal{D}$ from the subscript when context is clear, i.e., we may simply write $\Pr\{E\}, \mathbb{E}[f]$ for $\Pr_{\mathcal{D}}\{E\}, \mathbb{E}_{\mathcal{D}}[f]$.

In this paper, we denote halfspaces as a subset of $\mathbb{R}^d$ in the following way. For any $S_1, S_2 \subseteq \mathbb{R}^d$, we denote $S_1 \backslash S_2 = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} \in S_1, \mathbf{x} \notin S_2\}$ and $S^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} \notin S\}$. For any $t \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d$, let $l_t : \mathbb{R}^d \to \mathbb{R}$ be an affine function such that $l_t(\mathbf{x}, \mathbf{w}) = \langle \mathbf{x}, \mathbf{w}\rangle - t$. Then, a halfspace in $\mathbb{R}^d$ with threshold $t \in \mathbb{R}$ and normal vector $\mathbf{w}$ is defined as $h_t(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \mid l_t(\mathbf{x}, \mathbf{w}) \geq 0\}$ (resp. $h_t^c(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \mid l_t(\mathbf{x}, \mathbf{w}) \leq 0\}$). When a halfspace is homogeneous, we will drop the threshold from the subscript, i.e., when $t = 0$, we will write $h(\mathbf{w})$ instead of $h_0(\mathbf{w})$.

We will use an algorithm for robust list learning of sparse linear classifiers. Mossel & Sudan (2016) observed that the approach to robust regression for the sup norm used by Juba (2017) suffices (see Appendix A for an overview):

Theorem 2.1. *There is an algorithm for robust list-learning of linear classifiers with $s = O(1)$ nonzero coefficients from $m = O(\frac{1}{\alpha\epsilon}(s \log d + \log \frac{1}{\delta}))$ examples in polynomial time with list size $O((md)^s)$.*

3 Conditional Classification With Homogeneous Halfspaces

In this section, we present our algorithmic results for conditional classification with **homogeneous halfspaces (selectors)** on $\mathbb{R}^d$ for sparse linear classifiers or, more generally (cf. Theorem 2.1) **any**

small set of binary classifiers $\mathcal{C}$ under any distribution $\mathcal{D}$ with **standard normal** $\mathbf{x}$-marginals. For each classifier $c \in \mathcal{C}$, we find a homogeneous halfspace as the selector that minimizes its conditional classification loss, $\Pr\{c(\mathbf{x}) \neq y \mid \mathbf{x} \in h(\mathbf{w})\}$. Eventually, we choose the best classifier-selector pair as the output. Notice that, for homogeneous halfspaces under standard normal distributions, minimizing $\Pr\{c(\mathbf{x}) \neq y \mid \mathbf{x} \in h(\mathbf{w})\}$ is equivalent to minimizing $\Pr\{c(\mathbf{x}) \neq y \cap \mathbf{x} \in h(\mathbf{w})\}$ since every homogeneous halfspace $h(\mathbf{w})$ satisfies $\Pr_{\mathbf{x}\sim\mathcal{N}^d(0,1)}\{\mathbf{x} \in h(\mathbf{w})\} = 1/2$. Hence, we will only consider minimizing $\Pr\{c(\mathbf{x}) \neq y \cap \mathbf{x} \in h(\mathbf{w})\}$ in this section. The core challenge for our strategy is finding such a halfspace for each $c \in \mathcal{C}$. We give the details in the following sections.

3.1 Algorithm Overview

Algorithm 1: Conditional Classification With Homogeneous Halfspaces

1 **procedure** $\textsc{Cc}(\mathcal{D}, \mathcal{C}, \epsilon, \delta)$
2 $\quad T \leftarrow (4d + \ln(8|\mathcal{C}|/\delta))/\epsilon^4$
3 $\quad N \leftarrow 1600\ln^2(16T|\mathcal{C}|/\delta)/\epsilon^2$
4 $\quad \hat{\mathcal{D}} \leftarrow \ln(4|\mathcal{C}|T/\delta)/2\epsilon$ i.i.d. examples from $\mathcal{D}$
5 $\quad \mathbf{w}^{(0)} \leftarrow$ any basis
6 $\quad$ **for** $c \in \mathcal{C}$ **do**
7 $\quad\quad \mathcal{D}^{(c)} \leftarrow \mathcal{D}_{\mathbf{x}} \times \mathbb{1}\{c(\mathbf{x}) \neq y\}$
8 $\quad\quad \mathcal{W}^{(c)} \leftarrow \textsc{Psgd}\left(\mathcal{D}^{(c)}, T, N, \mathbf{w}^{(0)}\right) \cup \textsc{Psgd}\left(\mathcal{D}^{(c)}, T, N, -\mathbf{w}^{(0)}\right)$
9 $\quad\quad \mathbf{w}^{(c)} \leftarrow \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}^{(c)}} \Pr_{\hat{\mathcal{D}}}\{\mathbf{x} \in h(\mathbf{w}) \cap c(\mathbf{x}) \neq y\}$
10 $\quad$ **end**
11 $\quad$ **return** $\operatorname{argmin}_{\mathbf{w}^{(c)}} \Pr_{\hat{\mathcal{D}}}\{\mathbf{x} \in h(\mathbf{w}^{(c)}) \cap c(\mathbf{x}) \neq y\}$

In Algorithm 1, for each binary classifier $c \in \mathcal{C}$, we map the label y from $\mathcal{D}$ to $\mathbb{1}\{c(\mathbf{x}) \neq y\}$ to form a new distribution $\mathcal{D}^{(c)}$, then pass $\mathcal{D}^{(c)}$ to Algorithm 2 to obtain a sequence of halfspaces, and only keep the halfspace $h(\mathbf{w}^{(c)})$ with the smallest empirical conditional classification error for this classifier c. The last step picks out the classifier-selector pair that performs the best among all $c \in \mathcal{C}$ in terms of conditional classification error estimated on an large enough empirical distribution $\hat{\mathcal{D}}$.

Notably, the mapping step (line 7) for each $c \in \mathcal{C}$ essentially just creates another adversarial distribution $\mathcal{D}^{(c)}$, which is a key step to reduce the conditional classification problem to a **"one-sided" agnostic linear classification** problem. While directly optimizing over the conditional classification loss $\Pr\{\mathbf{x} \in h(\mathbf{w}) \cap c(\mathbf{x}) \neq y\}$ is intractable in general, it turns out that a simple convex surrogate approximation to the classification loss captures the "one-sided" nature for a standard normal distribution.

Algorithm 2: Projected SGD for $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$

1 **procedure** $\textsc{Psgd}(\mathcal{D}, T, N, \mathbf{w}^{(0)})$
2 $\quad \beta \leftarrow \sqrt{1/Td}$
3 $\quad$ **for** $i = 1, \ldots, T$ **do**
4 $\quad\quad \hat{\mathcal{D}}^{(i)} \leftarrow N$ i.i.d. samples from $\mathcal{D}$
5 $\quad\quad \mathbf{u}^{(i)} \leftarrow \mathbf{w}^{(i-1)} - \beta\mathbb{E}_{(\mathbf{x},y)\sim\hat{\mathcal{D}}^{(i)}}\left[g_{\mathbf{w}^{(i-1)}}(\mathbf{x}, y)\right]$
6 $\quad\quad \mathbf{w}^{(i)} \leftarrow \mathbf{u}^{(i)}/\|\mathbf{u}^{(i)}\|_2$
7 $\quad$ **end**
8 $\quad$ **return** $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)})$

Algorithm 2 is a variant of Stochastic Gradient Descent, and the loss function $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ we are minimizing is a **convex surrogate approximation** of the conditional classification error, known as ReLU. We formally define our loss function with respect to the distribution $\mathcal{D}$ to be $\mathcal{L}_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \cdot \max(0, \langle\mathbf{x}, \mathbf{w}\rangle)\right]$.

Inspired by Diakonikolas et al. (2020b), the updating policy in Algorithm 2 uses the projected gradient $g_{\mathbf{w}}(\mathbf{x}, y)$, defined as $g_{\mathbf{w}}(\mathbf{x}, y) = y \cdot \mathbf{x}_{\mathbf{w}^\perp} \cdot \mathbb{1}\{\mathbf{x} \in h(\mathbf{w})\}$. We will show in the next section that the goal of Algorithm 2 is not minimizing $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$, but the norm of the projected gradient $\|\mathbb{E}[g_{\mathbf{w}}]\|_2$.

Note that the objective function considered in Diakonikolas et al. (2020b) is completely different from ours so that their convergence analysis does not obviously hold for our surrogate loss $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$. Also, our choice of $g_{\mathbf{w}}(\mathbf{x}, y)$ is similar to that of Shen (2021). Nonetheless, the problem they were solving is agnostic linear classification and they used a quite different gradient descent policy.

3.2 Performance Guarantee

We introduce our main guarantee at first, but postpone the proof to Appendix C due to the page limit. As a sketch of the proof, we will see that Corollary 3.3 (an immediate result of Proposition 3.2) and Proposition 3.4 together indicate the optimality of Projected SGD, as captured by Lemma 3.5. Combined with a standard concentration analysis, this implies our main theorem.

Theorem 3.1 (Main Theorem)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{0, 1\}$ with standard normal $\mathbf{x}$-marginal, and $\mathcal{C}$ be a class of binary classifiers on $\mathbb{R}^d \times \{0, 1\}$. If there exists a unit vector $\mathbf{v} \in \mathbb{R}^d$ such that, for some sufficiently small $\epsilon \in [0, 1/e]$, $\min_{c \in \mathcal{C}} \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{v}) \cap c(\mathbf{x}) \neq y\} \leq \epsilon$, then, with at most $\tilde{O}(d/\epsilon^6)$ examples, Algorithm 1 will return a $\mathbf{w}^{(c)}$, with probability at least $1 - \delta$, such that $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left\{\mathbf{x} \in h(\mathbf{w}^{(c)}) \cap c(\mathbf{x}) \neq y\right\} = \tilde{O}(\sqrt{\epsilon})$ and run in time $O(d|\mathcal{C}|/\epsilon^6)$.*

The first and most important component that enables our approach is the following proposition, which simply says that, for any sub-optimal halfspace $h(\mathbf{w})$, the projected negative gradient $\mathbb{E}[-g_{\mathbf{w}}]$ of the surrogate loss $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ must have non-negligible projection on the normal vector of the optimal halfspace $h(\mathbf{v})$.

Proposition 3.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{0, 1\}$ with standard normal $\mathbf{x}$-marginal, and $g_{\mathbf{w}}(\mathbf{x}, y) = y \cdot \mathbf{x}_{\mathbf{w}^\perp} \mathbb{1}\{\mathbf{x} \in h(\mathbf{w})\}$. Suppose $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ are unit vectors such that $\theta(\mathbf{v}, \mathbf{w}) \in [0, \pi/2)$ and $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{v}) \cap y = 1\} \leq \epsilon$, then, if $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{w}) \cap y = 1\} \geq \frac{5}{2}(\epsilon\sqrt{\ln \epsilon^{-1}})^{1/2}$, there is $\langle \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[-g_{\mathbf{w}}(\mathbf{x}, y)], \bar{\mathbf{v}}_{\mathbf{w}^\perp} \rangle \geq \frac{2}{5}\epsilon\sqrt{\ln \epsilon^{-1}}$ for sufficiently small ϵ.*

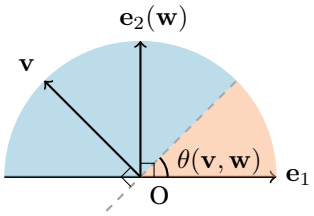


Figure 1: Blue area represents $h(\mathbf{v}) \cap h(\mathbf{w})$, orange area represents $h(\mathbf{w}) \backslash h(\mathbf{v})$.

We leave the formal proof to Appendix C due to the page limit. The proof is based on the following observation (also see Figure 1): When a homogeneous halfspace $h(\mathbf{w})$ is substantially sub-optimal, the probability of labels being true within the domain that the optimal halfspace $h(\mathbf{v})$ disagrees with it, i.e. $h(\mathbf{w}) \backslash h(\mathbf{v})$, must be large. However, the same probability cannot be too large in the optimal halfspace $h(\mathbf{v})$ and, hence, $h(\mathbf{v}) \cap h(\mathbf{w})$. Then, if the underlying distribution has a well-behaved $\mathbf{x}$-marginal, the l_2 norm of the expectation of $\mathbf{x}$ within that domain should also be large.

In fact, the observation also gives an insight into why we choose ReLU as the surrogate loss. As we are concerned about the one-sided loss, $\Pr\{\mathbf{x} \in h(\mathbf{w}) \cap y = 1\}$, we cannot make any assumption on the domain of $h^c(\mathbf{w})$, which is also the key difference between the analysis of agnostic classification and that of conditional classification. Notice that $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ completely "blocks" the information from $h^c(\mathbf{w})$ so that we only need to argue about $\mathbb{E}[g_{\mathbf{w}}(\mathbf{x}, y)]$ on the domain where we have control.

Besides, an important implication of Proposition 3.2 is that, once $\theta(\mathbf{v}, \mathbf{w}) \in [0, \pi/2)$ and $h(\mathbf{w})$ is sub-optimal, $\mathbb{E}[-g_{\mathbf{w}}(\mathbf{x}, y)]$ always "points" to $\mathbf{v}$. Then, the update step (line 5) in Algorithm 2 will make $\theta(\mathbf{v}, \mathbf{w})$ contractive, which will, in turn, guarantee that the assumption $\theta(\mathbf{v}, \mathbf{w}) \in [0, \pi/2)$ is satisfied in the next iteration. This property plays a key role in proving Lemma 3.5.

Negating the statement of Proposition 3.2 immediately gives the following corollary, which states that any approximate stationary point of $\mathcal{L}_{\mathcal{D}}(\bar{\mathbf{w}})$ admits an approximate optimal solution.

Corollary 3.3. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginal, and $g_{\mathbf{w}}(\mathbf{x},y) = y \cdot \mathbf{x}_{\mathbf{w}^\perp} \mathbb{1}\{\mathbf{x} \in h(\mathbf{w})\}$. Suppose $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ are unit vectors such that $\theta(\mathbf{v},\mathbf{w}) \in [0,\pi/2)$ and $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{v}) \cap y = 1\} \le \epsilon$, then, if a unit vector $\mathbf{w}$ satisfies that $\|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[g_{\mathbf{w}}(\mathbf{x},y)]\|_2 < \frac{2}{5}\epsilon\sqrt{\ln \epsilon^{-1}}$, there is $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{w}) \cap y = 1\} < \frac{5}{2}(\epsilon\sqrt{\ln \epsilon^{-1}})^{1/2}$ for sufficiently small ϵ.*

Proof. By Cauchy's inequality and our assumption, we have

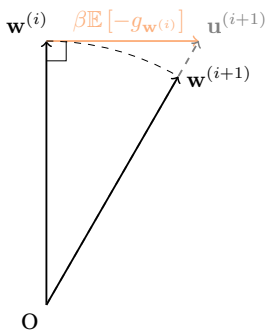
$$\left\langle \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[-g_{\mathbf{w}}(\mathbf{x},y)], \bar{\mathbf{v}}_{\mathbf{w}^\perp} \right\rangle \le \left\| \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[g_{\mathbf{w}}(\mathbf{x},y)] \right\|_2 < \frac{2}{5}\epsilon\sqrt{\ln \epsilon^{-1}}.$$

Then, negating the statement of Proposition 3.2 gives the desired result. □

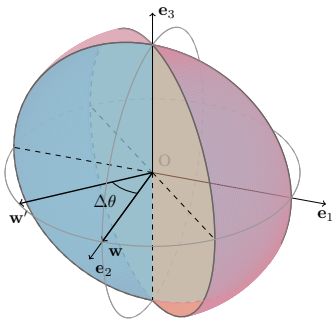
To effectively utilize Corollary 3.3, we also have to show that its assumption is satisfied. That is, at least one of the weight vectors, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$, produced by Algorithm 2 has small $\|\mathbb{E}[g_{\mathbf{w}}(\mathbf{x},y)]\|_2$. We show this can be achieved within a bounded number of iterations as the proposition below.

Proposition 3.4. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginal, $g_{\mathbf{w}}(\mathbf{x},y) = y \cdot \mathbf{x}_{\mathbf{w}^\perp} \cdot \mathbb{1}\{\mathbf{x} \in h(\mathbf{w})\}$, and $\mathcal{L}_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot \max(0, \langle \mathbf{x}, \mathbf{w}\rangle)]$. With $\beta = \sqrt{1/Td}$, after T iterations, the output $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)})$ in Algorithm 2 will satisfy $\mathbb{E}_{\hat{\mathcal{D}}^{(1)},\ldots,\hat{\mathcal{D}}^{(T)}\sim\mathcal{D}}[1/T\sum_{i=1}^{T}\|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[g_{\mathbf{w}^{(i)}}(\mathbf{x},y)]\|_2^2] \le \sqrt{d/T}$. In addition, if $T \ge (4d + \ln(1/\delta))/\epsilon^4$, then $\min_{i=1,\ldots,T}\|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[g_{\mathbf{w}^{(i)}}(\mathbf{x},y)]\|_2 \le \epsilon$ with probability at least $1-\delta$.*

We defer the formal proof to Appendix B. Our technique resembles the work of Diakonikolas et al. (2020b), which showed that, if the objective function is **bounded** and has **Lipschitz continuous** gradient, then the norm of its gradient converges in boundedly many iterations of (Projected) SGD.



(a) Weight update step (line 5) and projection step (line 6) in algorithm 2.



(b) Orange plane is the decision boundary of $h(\mathbf{w}')$, while blue plane is that of $h(\mathbf{w})$. $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ and $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w}')$ only differs in the two pink spherical sectors, which is dominated by $\Delta\theta$.

Figure 2: Boundedness of $\mathcal{L}_{\mathcal{D}}(\mathbf{w}^{(i)})$ and almost Lipschitz continuity of $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$.

However, the magnitude of $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ is dominated by $\|\mathbf{w}\|_2$, which could grow unbounded after many iterations, and its gradient $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ has a "jumping" point at zero, which is not Lipschitz continuous in general. So, the key to proving Proposition 3.4 is to overcome these issues.

On the one hand, the gradient update (line 5) of Algorithm 2 will always produce $\|\mathbf{w}^{(i)}\|_2 \ge \|\mathbf{w}^{(i-1)}\|_2$, while the projection step (line 6) of Algorithm 2 will always make $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ bounded, cf. Figure 2a.

On the other hand, it turns out that $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ is almost Lipschitz continuous under nice distributions such as a standard normal. Intuitively, if we perturb $\mathbf{w}$ a little bit to change it to $\mathbf{w}'$, it will only rotate the halfspace $h(\mathbf{w})$ by a very small angle, i.e. $\Delta\theta = \theta(\mathbf{w},\mathbf{w}')$ is small. And, it suffices to consider the difference between $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ and $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w}')$ on a 3-dimensional subspace as shown

in figure 2b. Now, if the density of distribution $\mathcal{D}$ is not concentrated too much in any small spherical sectors in the subspace, it implies that the change of $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ is dominated by $\Delta\theta$ (see Figure 2b), which is insignificant. This observation indicates that $\nabla_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ is Lipschitz continuous under anti-concentrated distributions unless $\|\mathbf{w}\|_2$ is extremely small.

Given Corollary 3.3 and Proposition 3.4, we show that in the list of parameters returned by Algorithm 2, at least one of them is approximately optimal:

Lemma 3.5. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginal, and $g_{\mathbf{w}}(\mathbf{x}, y) = y \cdot \mathbf{x}_{\mathbf{w}^\perp} \cdot \mathbb{1}\{\mathbf{x} \in h(\mathbf{w})\}$. Suppose $\mathbf{v} \in \mathbb{R}^d$ is a unit vectors such that $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{v}) \cap y = 1\} \leq \epsilon$, if $T \geq (4d + \ln(2/\delta))/\epsilon^4$, $N \geq 1600\ln^2(4T/\delta)/\epsilon^2$, and $\theta(\mathbf{v}, \mathbf{w}^{(0)}) \in [0, \pi/2)$, at least one of $\mathbf{w} \in \mathcal{W} = \{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}\}$ returned by Algorithm 2 satisfies*

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\{\mathbf{x} \in h(\mathbf{w}) \cap y = 1\} \leq \frac{5}{2}(\epsilon\sqrt{\ln \epsilon^{-1}})^{1/2}$$

with probability at least $1 - \delta$ for some sufficiently small $\epsilon \in [0, 1/e]$.

We defer the formal proof to Appendix C, but sketch the idea here. Observe that combining the results of Corollary 3.3 and Proposition 3.4 already yields Lemma 3.5. So, all we need to do is make sure that the assumption $\theta(\mathbf{v}, \mathbf{w}) \in [0, \pi/2)$ in Corollary 3.3 is satisfied.

Notice that, in the sequence of parameters $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$ returned by Algorithm 2, every $\mathbf{w}^{(i)}$ must be significantly sub-optimal until we see a $\mathbf{w}$ such that $\Pr\{\mathbf{x} \in h(\mathbf{w}) \cap y = 1\} \leq \frac{5}{2}(\epsilon\sqrt{\ln \epsilon^{-1}})^{1/2}$. If such a sub-optimal halfspace $h(\mathbf{w}^{(i)})$ also satisfies $\theta(\mathbf{v}, \mathbf{w}^{(i)}) \in [0, \pi/2)$, its projected negative gradient $\mathbb{E}[-g_{\mathbf{w}^{(i)}}]$ must has positive projection on $\bar{\mathbf{v}}_{\mathbf{w}^\perp}$ by Proposition 3.2. Using such a $\mathbb{E}[-g_{\mathbf{w}^{(i)}}]$ to update $\mathbf{w}^{(i)}$ in Algorithm 2 will always produce $\theta(\mathbf{v}, \mathbf{w}^{(i+1)}) \leq \theta(\mathbf{v}, \mathbf{w}^{(i)})$. Thus, by an inductive argument, we can show that the first $\mathbf{w}^{(t)}$ such that $\|\mathbb{E}[g_{\mathbf{w}^{(t)}}]\|_2 < \frac{2}{5}\epsilon\sqrt{\ln \epsilon^{-1}}$ must satisfy $\theta(\mathbf{v}, \mathbf{w}^{(t)}) \in [0, \pi/2)$, which enables the application of Corollary 3.3.

4 Conditional Classification With General Halfspaces Is Hard

In this section, we show that it is computationally hard to obtain a small additive error for conditional classification with general halfspaces for any finite class of classifiers $\mathcal{C}$, even under distributions with standard normal $\mathbf{x}$-marginals. Specifically, we show that, for each classifier $c \in \mathcal{C}$, approximating the optimal conditional classification loss over the class of general halfspaces on $\mathbb{R}^d$ with an **additive error** is at least as hard as achieving the same additive error for agnostic linear classification, which is known to be computationally hard (Diakonikolas et al., 2023). Further, we show that any $(1+\alpha)$-approximation algorithm for conditional classification implies an $(1+\alpha)$-approximation algorithm for standard classification, down to polynomially small losses. (The converse is not known to hold.)

The hardness of distribution-specific conditional classification is based on the sub-exponential hardness of "continuous Learning With Errors" (cLWE), which is a variant of the "Learning With Errors" (LWE) assumption. Informally speaking, in the problem of LWE, we are given labelled examples from two hypothesis cases. In one case, the labels are biased by some secret vector, while, in another case, the labels are generated uniformly at random. We wish to distinguish between these cases. We formally define the problem of LWE (Regev, 2009), following Diakonikolas et al. (2023):

Definition 4.1 (Learning With Errors)**.** *For $m, d \in \mathbb{N}$, $q \in \mathbb{R}_+$, let $\mathcal{D}_{sample}, \mathcal{D}_{secret}, \mathcal{D}_{noise}$ be distributions on $\mathbb{R}^d, \mathbb{R}^d, \mathbb{R}$ respectively. In the LWE$(m, \mathcal{D}_{sample}, \mathcal{D}_{secret}, \mathcal{D}_{noise}, \mathrm{mod}_q)$ problem, with m independent samples $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$, we want to distinguish between the following two cases:*

- ***Alternative hypothesis**: each $(\mathbf{x}^{(i)}, y^{(i)})$ is generated as $y^{(i)} = \mathrm{mod}_q(\langle \mathbf{x}^{(i)}, \mathbf{s}\rangle + z)$, where $\mathbf{x}^{(i)} \sim \mathcal{D}_{sample}, \mathbf{s} \sim \mathcal{D}_{secret}, z \sim \mathcal{D}_{noise}$.*
- ***Null hypothesis**: each $y^{(i)}$ is sampled uniformly at random on the support of its marginal distribution in the alternative hypothesis, independent of $\mathbf{x}^{(i)} \sim \mathcal{D}_{sample}$.*

An algorithm is said to be able to *solve the LWE problem with Δ advantage* if the probability that the algorithm outputs "alternative hypothesis" is Δ larger than the probability that it outputs "null hypothesis" when the given data is sampled from the alternative hypothesis distribution.

Let $\mathbb{S}^{d-1} := \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\right\}$, $\mathbb{R}_q := [0, q)$, and $\mathrm{mod}_q : \mathbb{R}^d \to \mathbb{R}_q^d$ to be the function that applies mod_q operation on each coordinate of $\mathbf{x}$. Essentially, the hardness of cLWE is based on the sub-exponential hardness of LWE (see Appendix D). We formally state the assumption of sub-exponential hardness of cLWE as follows.

Assumption 4.2 ((Gupte et al., 2022; Diakonikolas et al., 2023) Sub-exponential cLWE Assumption)**.** *For any $d \in \mathbb{N}$, any constants $\kappa \in \mathbb{N}, \alpha \in (0,1), \beta \in \mathbb{R}_+$ and any $\log^\beta d \le k \le Cd$ where $C > 0$ is a sufficiently small universal constant, the problem LWE$(d^{O(k^\alpha)}, \mathcal{N}^d(0,1), \mathbb{S}^{d-1}, \mathcal{N}(0,\sigma^2), \mathrm{mod}_T)$ over $\mathbb{R}^d$ with $\sigma \ge k^{-\kappa}$ and $T = 1/C'\sqrt{k \log d}$, where $C' > 0$ is a sufficiently large universal constant, cannot be solved in time $d^{O(k^\alpha)}$ with $d^{-O(k^\alpha)}$ advantage.*

For simplicity, we define $y \equiv \mathbb{1}\{c(\mathbf{x}) \neq y'\}$ for $(\mathbf{x}, y') \sim \mathcal{D}'$ and construct the distribution $(\mathbf{x}, y) \sim \mathcal{D}$. Notice that, in agnostic settings, since $\mathcal{D}'$ is worst case, $\mathcal{D}$ is also worst case. Therefore, this replacement does not affect the difficulty of the problems we consider.

Normally, for the problem of agnostic classification, one would consider its loss function to be the expected disagreement between the classifier and the labelling. However, it is more convenient for us to consider a labelling $y = 1$ as an "occurrence of an error" and, hence, define the loss function in terms of agreement to compare with the conditional classification loss. Specifically, for any binary classifier as a subset $S \subseteq \mathbb{R}^d$ and any distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{0,1\}$, we define the classification loss:

$$\mathrm{err}_{\mathcal{D}}(S) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left\{y = \mathbb{1}\{\mathbf{x} \in S\}\right\}. \tag{1}$$

Note that this definition of classification loss is essentially the same as the traditional one defined in terms of disagreement since we can convert from one to another by simply negating the labelling.

Analogously, for any binary classifiers as subsets $S, T \subseteq \mathbb{R}^d$ and any distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{0,1\}$, we denote the conditional classification loss by

$$\mathrm{err}_{\mathcal{D}|T}(S) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left\{y = \mathbb{1}\{\mathbf{x} \in S\} \mid \mathbf{x} \in T\right\}. \tag{2}$$

For simplicity, we write $\mathrm{err}_{\mathcal{D}|T}$ instead of $\mathrm{err}_{\mathcal{D}|T}(S)$ when $S \equiv T$.

We state our distribution-specific hardness result for conditional classification as Theorem 4.3.

Theorem 4.3 (Hardness Of Conditional Classification)**.** *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginals, $\mathcal{H}$ be the class of halfspaces on $\mathbb{R}^d$, and define $\mathcal{H}_{\mathcal{D}}^{a,b} = \{h_t(\mathbf{w}) \in \mathcal{H} \mid \Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{\mathbf{x} \in h_t(\mathbf{w})\} \in [a,b]\}$ for any $0 \le a \le b \le 1$. Under Assumption 4.2, for any constant $\alpha \in (0,2)$, $\gamma > 1/2$ and any $c/\sqrt{d \log d} \ge \epsilon \le 1/\log^\gamma d$ where c is a sufficiently large constant, there is no algorithm that can find a halfspace $h_{t'}(\mathbf{w}) \in \mathcal{H}_{\mathcal{D}}^{a,b}$ such that $\mathrm{err}_{\mathcal{D}|h_{t'}(\mathbf{w})} \le \min_{h_t(\mathbf{u}) \in \mathcal{H}_{\mathcal{D}}^{a,b}} \mathrm{err}_{\mathcal{D}|h_t(\mathbf{u})} + \epsilon$ and runs in time $d^{O(1/(\epsilon\sqrt{\log d})^\alpha)}$.*

Theorem 4.3 is actually a simple consequence of Proposition 4.5 and Lemma 4.6, where the former one shows that conditional classification is at least as hard as agnostic classification and the latter one states the hardness of agnostically learning halfspaces.

Our main contribution is Proposition 4.5, but before getting into it, we first show a simple but critical observation that reveals the relationship between $\mathrm{err}_{\mathcal{D}}(S)$ and $\mathrm{err}_{\mathcal{D}|S}$. That is, the loss of agnostic classification can be explictly expressed by the loss of conditional classification.

Lemma 4.4 (Classification Error Decomposition)**.** *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$ and S be any subset of $\mathbb{R}^d$, there are $\mathrm{err}_{\mathcal{D}}(S) = 2\mathrm{err}_{\mathcal{D}|S}\Pr_{\mathcal{D}}\{\mathbf{x} \in S\} + \Pr_{\mathcal{D}}\{y = 0\} - \Pr_{\mathcal{D}}\{\mathbf{x} \in S\}$ as well as $\mathrm{err}_{\mathcal{D}}(S) = 2\mathrm{err}_{\mathcal{D}|S^c}(S)\Pr_{\mathcal{D}}\{\mathbf{x} \in S^c\} + \Pr_{\mathcal{D}}\{y = 1\} - \Pr_{\mathcal{D}}\{\mathbf{x} \in S^c\}$.*

Due to page limits, we defer its proof to Appendix D. Lemma 4.4 is a powerful result since it allows us to establish a reduction from classification to conditional classification.

Briefly speaking, if we know $\Pr\{\mathbf{x} \in S^*\}$ for some optimal solution S^* to the agnostic classification problem, we can approximate $\mathrm{err}_{\mathcal{D}}(S^*)$ by approximating its conditional classification loss, i.e. $\mathrm{err}_{\mathcal{D}|S^*}$. Even though we do not know $\Pr\{\mathbf{x} \in S^*\}$, we can guess a small range containing $\Pr\{\mathbf{x} \in S^*\}$, and enforce such a constraint just as in Definition 1.1. Then, we sweep over all such small intervals and one of the instances being solved must include $\Pr\{\mathbf{x} \in S\}$. Once we take these intervals small enough, it won't incur a significant error. We use this strategy to prove Proposition 4.5, but the formal proof is deferred to Appendix D due to the page limit.

Proposition 4.5 (Reduction In Additive Form). *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$, $\mathcal{H}$ be any subset of the power set of $\mathbb{R}^d$ closed under complement, and define $\mathcal{H}_{\mathcal{D}}^{a,b} = \{S \in \mathcal{H} \mid \Pr_{\mathcal{D}}\{\mathbf{x} \in S\} \in [a,b]\}$ for any $0 \le a \le b \le 1$. For any such a, b and $\epsilon, \delta > 0$, given sample access to $\mathcal{D}$, if there exists an algorithm $\mathcal{A}_1(\epsilon, \delta, a, b)$ running in time $\mathrm{poly}(d, 1/\epsilon, 1/\delta)$, that outputs $S_1 \in \mathcal{H}_{\mathcal{D}}^{a,b}$ such that $\mathrm{err}_{\mathcal{D}|S_1} \le \min_{S \in \mathcal{H}_{\mathcal{D}}^{a,b}} \mathrm{err}_{\mathcal{D}|S} + \epsilon$ with probability as least $1-\delta$, there exists another algorithm $\mathcal{A}_2(\epsilon, \delta)$, that runs in time $\mathrm{poly}(d, 1/\epsilon, 1/\delta)$, and outputs $S_2 \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(S_2) \le \min_{S \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(S) + 6\epsilon$ with probability at least $1-\delta$.*

Furthermore, the following distribution-specific hardness result states that agnostically learning halfspaces up to small additive error is computationally hard.

Lemma 4.6 (Corollary 3.2 of Diakonikolas et al. (2023)). *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$ with standard normal $\mathbf{x}$-marginals, and $\mathcal{H}$ be the class of halfspaces on $\mathbb{R}^d$. Under Assumption 4.2, for any constant $\alpha \in (0,2)$, $\gamma > 1/2$ and any $c/\sqrt{d \log d} \ge \epsilon \le 1/\log^{\gamma} d$ where c is a sufficiently large constant, there is no algorithm that can find a halfspace $h_{t'}(\mathbf{v}) \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h_{t'}(\mathbf{v})) \le \min_{h_t(\mathbf{u}) \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h_t(\mathbf{u})) + \epsilon$ and runs in time $d^{O(1/(\epsilon\sqrt{\log d})^{\alpha})}$.*

Since Proposition 4.5 holds for halfspaces on $\mathbb{R}^d$, conditional learning has at least the same hardness by combining Proposition 4.5 and Lemma 4.6.

Analogously, a reduction in multiplicative form can also be obtained using a similar analysis to that in the proof of Proposition 4.5. In particular, we show that if there exists a multiplicative approximation algorithm for conditional classification with factor $1+\alpha$, there must exist another multiplicative approximation algorithm for classification in agnostic setting with the same factor $1+\alpha$.

Claim 4.7 (Reduction In Multiplicative Form). *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{0,1\}$, $\mathcal{H}$ be any subset of the power set of $\mathbb{R}^d$ closed under complement, and define $\mathcal{H}_{\mathcal{D}}^{a,b} = \{S \in \mathcal{H} \mid \Pr_{\mathcal{D}}\{\mathbf{x} \in S\} \in [a,b]\}$ for any $0 \le a \le b \le 1$. If there exists an algorithm $\mathcal{A}_1(\alpha, \delta, a, b)$ that given sample access to $\mathcal{D}$, any such a, b, and $\alpha, \epsilon, \delta > 0$, runs in time $\mathrm{poly}(d, 1/\alpha, 1/\delta)$, and outputs $S_1 \in \mathcal{H}_{\mathcal{D}}^{a,b}$ such that $\mathrm{err}_{\mathcal{D}|S_1} \le (1+\alpha)\min_{S \in \mathcal{H}_{\mathcal{D}}^{a,b}} \mathrm{err}_{\mathcal{D}|S}$ with probability as least $1-\delta$, there exists another algorithm $\mathcal{A}_2(\alpha, \epsilon, \delta)$ that runs in time $\mathrm{poly}(d, 1/\alpha, 1/\epsilon, 1/\delta)$, and outputs $S_2 \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(S_2) \le (1+\alpha)(\min_{S \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(S) + 4\epsilon)$ with probability at least $1-\delta$.*

Again, we defer the proof to Appendix D because of page limits. Although there is an extra 4ϵ additive error in the final guarantee of Claim 4.7, we can afford to take ϵ polynomially small w.r.t. d, α, δ, thus obtaining the multiplicative error guarantee down to polynomially small error. Informally we observe that Proposition 4.5 and Claim 4.7 indicate that any form of approximation algorithm for conditional classification yields an approximation algorithm of the same factor for agnostic classification. In the case of multiplicative approximation in particular, the reverse is not known and we observe that it might be strictly harder to approximate the conditional classification objective.

5 Limitations And Future Work

Our algorithmic result is limited in three aspects. First and foremost, the restriction of selectors to homogeneous halfspaces is a major drawback especially for the task of conditional classification. Indeed, the advantage of conditional classification with halfspaces compared with regular linear classification really shines when we have the ability to select a minority of the data distribution. Therefore, even with guarantees worse than $\tilde{O}(\sqrt{\epsilon})$, moving from homogeneous halfspaces to general halfspaces would constitute a significant advace. Another limitation of our result is the strong assumption on the marginal distribution. Real-world data almost never has standard normal marginals, and testing for a standard normal distribution is costly. Hence, it's worth trying to extend our result to more general classes of distributions, such as log-concave distributions. Last but not the least, one can also try to improve our error guarantee under the current setting as the error guarantee $O(\sqrt{\epsilon})$ appears sub-optimal.

References

Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms*

(SODA), pp. 1279–1297. SIAM, 2021.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Diego Calderon, Brendan Juba, Sirui Li, Zongyi Li, and Lisa Ruan. Conditional linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR, 2020.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 67–82. Springer, 2016.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. A polynomial time algorithm for learning halfspaces with tsybakov noise. *arXiv preprint arXiv:2010.01705*, 2020a.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pp. 1486–1513. PMLR, 2020b.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33: 18540–18549, 2020c.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pp. 5118–5141. PMLR, 2022.

Ilias Diakonikolas, Daniel Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *International Conference on Machine Learning*, pp. 7922–7938. PMLR, 2023.

Ilias Diakonikolas, Daniel Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 36, 2024.

Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pp. 3417–3426. PMLR, 2021.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2179–2187. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/gangrade21a.html`.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Roei Gelbhart and Ran El-Yaniv. The relationship between agnostic selective classification, active learning and the disagreement coefficient. *Journal of Machine Learning Research*, 20(33):1–38, 2019. URL http://jmlr.org/papers/v20/17-147.html.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1162–1173. IEEE, 2022.

John Hainline, Brendan Juba, Hai S. Le, and David Woodruff. Conditional sparse l_p-norm regression with optimal probability. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1042–1050. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/hainline19a.html.

Blaise Hanczar and Edward R Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2008.

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

David Haussler. Quantifying inductive bias: Ai learning algorithms and valiant's learning framework. *Artificial intelligence*, 36(2):177–221, 1988.

Daniel Hsu, Jizhou Huang, and Brendan Juba. Distribution-specific auditing for subgroup fairness. In *5th Symposium on Foundations of Responsible Computing (FORC 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024.

Brendan Juba. Learning abductive reasoning using random examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Feb. 2016. doi: 10.1609/aaai.v30i1.10099. URL https://ojs.aaai.org/index.php/AAAI/article/view/10099.

Brendan Juba. Conditional sparse linear regression. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2017.

Brendan Juba, Zongyi Li, and Evan Miller. Learning abduction under partial observability. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1035–1046, 2018.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in neural information processing systems*, 36, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, pp. 4753–4761. PMLR, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In Claire Vernade and Daniel Hsu (eds.), *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 822–867. PMLR, 25–28 Feb 2024c. URL https://proceedings.mlr.press/v237/mao24a.html.

Elchanan Mossel and Madhu Sudan. Personal communication, 2016.

Andrea Pugnana and Salvatore Ruggieri. Auc-based selective classification. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2494–2514. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/pugnana23a.html.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM (JACM)*, 56(6):1–40, 2009.

Jie Shen. On the power of localized perceptron for label-optimal learning of halfspaces with adversarial noise. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9503–9514. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/shen21a.html.

Yair Wiener and Ran El-Yaniv. Agnostic selective classification. *Advances in neural information processing systems*, 24, 2011.

Yair Wiener and Ran El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.

Mengxue Zhang, Tushar Mathew, and Brendan Juba. An improved algorithm for learning to perform exception-tolerant abduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.