Metric-Fair Active Learning

Jie Shen [1] **Nan Cui** [1] **Jing Wang** [2]

Abstract

Active learning has become a prevalent technique for designing label-efficient algorithms, where the central principle is to only query and fit "informative" labeled instances. It is, however, known that an active learning algorithm may incur unfairness due to such instance selection procedure. In this paper, we henceforth study metric-fair active learning of homogeneous halfspaces, and show that under the distribution-dependent PAC learning model, fairness and label efficiency can be achieved simultaneously. We further propose two extensions of our main results: 1) we show that it is possible to make the algorithm robust to the adversarial noise – one of the most challenging noise models in learning theory; and 2) it is possible to significantly improve the label complexity when the underlying halfspace is sparse.

1. Introduction

Deep learning has become the driving force behind modern artificial intelligence. However, it requires massive amount of labeled data for model training. Though there is a massive amount of unlabeled data available in many applications, the labels are typically precious and expensive to acquire, especially in the areas of medicine and physiology. In this regard, active learning was broadly utilized as a paradigm to learn a good model with significantly fewer labels by designing strategies to adaptively select informative instances to annotate (Cohn et al., 1994; Dasgupta et al., 2005; Balcan et al., 2006; Dasgupta, 2009).

On the other hand, recently practitioners from different disciplines highlighted the ethical and legal challenges posed by machine learning systems which are with potential to discriminate against specific population groups (Chouldechova & Roth, 2020). The rising concern is that the designed algorithms achieve appealing prediction accuracy, yet cannot recognize ethical or moral feelings, and they will likely output a solution that treats vulnerable groups unfairly. For instance, Apple's credit card has been investigated by financial regulators after customers discovered that the lending algorithms were discriminating against women. In this light, there is a growing interest in incorporating different fairness criteria into algorithmic design, aiming to guarantee that similar individuals or groups will be treated equally (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Yona & Rothblum, 2018; Liu et al., 2018; Dwork & Ilvento, 2019; Liu et al., 2019; Dwork et al., 2020a;b; Ding et al., 2021).

In this paper, we study the two properties that seemingly are odd with each other, and propose the first provable active learning algorithm to address the general concern that the instance selection paradigm in active learning may lead to unfairness. Our goal is three-fold: 1) designing a computationally efficient algorithm that learns the underlying hypothesis class under the probably approximately correct (PAC) model of Valiant (1984); 2) significantly reducing the label complexity using active learning techniques; and 3) ensuring fairness on the unseen data, i.e. generalization ability of the fairness guarantee. The first two objectives have been broadly studied in the literature and were achieved by many active learning algorithms (Balcan et al., 2007; Awasthi et al., 2017); hence our main contribution falls into the design of a new family of active learning algorithms that additionally satisfy the fairness guarantee.

1.1. Formal setup

We study efficient PAC learning of homogeneous halfspaces (Valiant, 1984), which is arguably one of the most important problems in learning theory (Rosenblatt, 1958). Denote by $\mathcal{X} := \mathbb{R}^d$ the instance space and by $\mathcal{Y} := \{-1, 1\}$ the label space. The class of homogeneous halfspaces is given by $\mathcal{H} := \{x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, \|w\|_2 = 1\}$. Let D be the joint distribution on $\mathcal{X} \times \mathcal{Y}$ and denote by D_X the marginal distribution on $\mathcal{X}$. For any hypothesis $w \in \mathcal{H}$, we define the error rate as $\text{err}_D(w) := \Pr_{(x,y)\sim D}\big(\text{sign}(w \cdot x) \neq y\big)$.

Let EX_D be the sample generation oracle such that each time the learner makes a call, it returns a labeled instance

[1]Department of Computer Science, Stevens Institute of Technology, Hoboken, New Jersey, USA. [2]Amazon, New York City, New York, USA. Correspondence to: Jie Shen <jie.shen@stevens.edu>, Nan Cui <ncui@stevens.edu>, Jing Wang <jing.julia.wang@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

(x, y) that is randomly drawn from D. In the active learning setting, however, each time EX_D is called, a labeled instance (x, y) is still randomly drawn from D, but the oracle only returns the instance x. The learner must make another call to a label revealing oracle EX_D^Y to obtain the label y. Let $\epsilon \in (0, 1)$ be the target classification error rate, and $\delta \in (0, 1)$ be the failure probability. We say $\mathcal{H}$ is PAC learnable if there exists a learning algorithm $\mathcal{A}$, quantities $n^{\mathcal{A}}_{\epsilon,\delta}$ and $m^{\mathcal{A}}_{\epsilon,\delta}$ satisfying the following: given any ϵ and δ, by making $n^{\mathcal{A}}_{\epsilon,\delta}$ calls to EX_D and $m^{\mathcal{A}}_{\epsilon,\delta}$ calls to EX_D^Y, $\mathcal{A}$ outputs a halfspace $\hat{w}$ with $\mathrm{err}_D(\hat{w}) \le \min_{w\in\mathcal{H}} \mathrm{err}_D(w) + \epsilon$ with probability at least $1 - \delta$ (over the draw of the data and the internal randomness of $\mathcal{A}$). The minimum of $n^{\mathcal{A}}_{\epsilon,\delta}$ and $m^{\mathcal{A}}_{\epsilon,\delta}$ over all possible algorithms are termed the *sample complexity* and *label complexity*, respectively.

In addition to the PAC guarantee, we also aim to establish fairness guarantee. We consider the notion of approximate metric-fairness due to Yona & Rothblum (2018), yet with a slight modification.

Definition 1 (Approximate metric-fairness)**.** Given a metric $\zeta : \mathcal{X} \times \mathcal{X} \to [0, 1]$, let the fairness error be

$$f_\zeta(w; D_X) := \Pr_{D_X \times D_X}\big(\big|w \cdot x - w \cdot x'\big| > \zeta(x, x')\big).$$

A hypothesis w is said to be α-approximately metric-fair if $f_\zeta(w; D_X) \le \alpha$. We call α the fairness error rate.

We note that when $\alpha = 0$, Definition 1 reduces to the notion of *perfect* metric-fairness (Dwork et al., 2012). That is, for all $(x, x') \in \mathcal{X} \times \mathcal{X}$,

$$\big|w \cdot x - w \cdot x'\big| \le \zeta(x, x') \text{ almost surely.}$$

Yet, as shown in Yona & Rothblum (2018), even a perfectly metric-fair hypothesis exists and has zero error rate, for some simple learning problem, it cannot be found in polynomial time by any perfectly metric-fair algorithm. Therefore, throughout the paper, we will only consider finding a hypothesis with the property of approximate metric-fairness, which relaxes the perfectness in such a way that for an α fraction of the pairs, the metric-fairness property may not hold. It is also worth mentioning that Yona & Rothblum (2018) considered a slightly more general definition where w is said metric-fair if $\big|w \cdot x - w \cdot x'\big| \le \zeta(x, x') + \gamma$ for some slack parameter $\gamma \ge 0$. We find such relaxation seems unnecessary and our analysis will be different in the way that we design a computationally efficient learning algorithm. Specifically, we will utilize a different fairness loss function; see Section 3.

Now we are in the position to state the probably approximately correct and fair (PACF) learning problem.

Definition 2 (PACF learning)**.** A learning algorithm PACF-learns a hypothesis class $\mathcal{H}$ if for any underlying distribution D, target classification error rate $\epsilon \in (0, 1)$, confidence $\delta \in (0, 1)$, fairness metric $\zeta(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to [0, 1]$, fairness error rate $\alpha \in (0, 1)$, it randomly draws a number of samples from D and with probability $1 - \delta$ over the draw, outputs a halfspace $\hat{w}$ satisfying: 1) $\mathrm{err}_D(\hat{w}) \le \min_{w\in\mathcal{H}^\alpha} \mathrm{err}_D(w) + \epsilon$; and 2) $f_\zeta(\hat{w}; D_X) \le \alpha$, where $\mathcal{H}^\alpha \subset \mathcal{H}$ consists of all halfspaces that are α-approximately metric-fair.

It is worth mentioning that in terms of classification error, Yona & Rothblum (2018) competed with the best hypothesis in a subclass $\mathcal{H}^{\alpha-\epsilon_\alpha}$ where $\epsilon_\alpha \in (0, \alpha)$ (but the most interesting regime is $\epsilon_\alpha = \Theta(\alpha)$), known as *relaxed* PACF learning. In this work, we alternatively make a realizable assumption that there exists a perfectly metric-fair target halfspace w^* in $\mathcal{H}^\alpha$ to avoid the relaxation of PACF learnability. This naturally interpolates the settings in Dwork et al. (2012) and Yona & Rothblum (2018) and thus our results can be thought of as circumventing the computational hardness of finding the perfectly metric-fair hypothesis while permitting simpler technical analysis.

1.2. Main results

The main contribution of this paper is a metric-fair active learning algorithm that fortifies state-of-the-art active learning algorithms with the PACF guarantee. In this section, we summarize our main results; readers are referred to Section 5 for a more comprehensive discussion.

We will present a basic PACF algorithm with label efficiency. Then we will show how to improve this basic algorithm so that it can tolerate the adversarial label noise (Kearns et al., 1992) and can learn sparse halfspaces (Littlestone, 1987).

All of our analysis hinges on a mild assumption on the marginal distribution D_X.

Assumption 1. The marginal distribution D_X is isotropic log-concave on $\mathcal{X}$; namely, it has zero mean, unit covariance matrix, and the logarithm of its density function is concave.

Observe that the family of isotropic log-concave distributions is fairly standard and general (Lovász & Vempala, 2007b; Vempala, 2010; Balcan & Long, 2013). Without any assumptions on the distribution, active learning may fail to provide any improvement over passive learning in terms of label efficiency; see an example given by Dasgupta (2005).

Our first result concerns label-efficient PACF learning in the noise-free setting, where there is a perfect hypothesis $w^* \in \mathcal{H}$ that incurs zero error rate.

Theorem 3. *If Assumption 1 is satisfied and there exists $w^* \in \mathcal{H}$ with $\mathrm{err}_D(w^*) = 0$ and $f_\zeta(w^*; D_X) = 0$, then there is an efficient algorithm that PACF learns $\mathcal{H}$. In addition, the label complexity is $O\big(d \cdot \mathrm{polylog}(\frac{1}{\epsilon}, \frac{1}{\alpha})\big)$.*

Observe that the obtained label complexity significantly im-

Table 1. **Comparison to most relevant works.** Compared to Zhang (2018), our algorithms can produce metric-fair hypotheses. Compared to Yona & Rothblum (2018), we have exponential improvement on the dependence of ϵ and α, and are robust to adversarial noise. We remark that the PACF learnability of halfspaces was not set out in Yona & Rothblum (2018) but their results implied what we list here.

Work	Label Complexity	Metric Fairness	Noise Tolerance
Zhang (2018)	$\text{polylog}(\frac{1}{\epsilon}) \cdot O(t \cdot \text{polylog}(d))$	✗	✓
Yona & Rothblum (2018)	$\frac{1}{\epsilon^2} \cdot \frac{1}{\alpha^2} \cdot O(t \cdot \text{polylog}(d))$	✓	✗
This Work (Theorem 4)	$\text{polylog}(\frac{1}{\epsilon}, \frac{1}{\alpha}) \cdot O(d)$	✓	✓
This Work (Theorem 5)	$\text{polylog}(\frac{1}{\epsilon}, \frac{1}{\alpha}) \cdot O(t \cdot \text{polylog}(d))$	✓	✓

proves upon the one of Yona & Rothblum (2018): theirs is proportional to $\frac{1}{(\epsilon\alpha)^2}$ while we have a poly-logarithmic dependence on both $\frac{1}{\epsilon}$ and $\frac{1}{\alpha}$. This is due to our new algorithmic design and the distributional assumption we made.

We then consider a more challenging setting where no perfect halfspace exists in $\mathcal{H}$ with zero error rate, known as the adversarial noise (Haussler, 1992; Kearns et al., 1992). Note that this is a very challenging label noise and only recently have efficient algorithms been established, though without fairness guarantees (Awasthi et al., 2017; Yan & Zhang, 2017; Shen, 2021a).

Assumption 2. The distribution D is said to satisfy the η-adversarial-noise condition if there is a halfspace $w^* \in \mathcal{H}$ with $\text{err}_D(w^*) \leq \eta$.

Theorem 4. *If Assumptions 1 and 2 are satisfied, and $f_\zeta(w^*; D_X) = 0$, then there is a polynomial-time algorithm that PACF learns $\mathcal{H}$ if $\eta \leq O(\epsilon)$. In addition, the label complexity is $O\big(d \cdot \text{polylog}(\frac{1}{\epsilon}, \frac{1}{\alpha})\big)$.*

Lastly, the margin-based active learning framework allows us to explore learning of structured halfspaces. In particular, we are interested in learning of t-sparse halfspaces and the goal is to obtain label complexity that is sublinear in the dimension d, a property termed attribute-efficiency (Littlestone, 1987). Such property was also broadly studied in statistics and signal processing communities (Chen et al., 1998; Tibshirani, 1996; Candès & Tao, 2005).

Assumption 3. The hypothesis class consists of s-sparse halfspaces, i.e. $\mathcal{H} = \{x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, \|w\|_2 = 1, \|w\|_0 \leq t\}$, where $\|w\|_0$ counts the number of non-zero elements of w.

Theorem 5 (Theorem 9, informal)**.** *If Assumptions 1, 2 and 3 are satisfied and $f_\zeta(w^*; D_X) = 0$, then there is a polynomial-time algorithm that PACF learns $\mathcal{H}$ if $\eta \leq O(\epsilon)$. In addition, the label complexity is $O\big(t \cdot \text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\alpha})\big)$.*

Observe that both Theorem 3 and Theorem 4 are just special cases of the above result. Without the fairness constraint, the setting of Theorem 5 has been studied in Zhang (2018). In fact, our high-level idea of algorithmic design is inspired by that work. The crucial difference lies in the incorporation of the metric-fairness and hence, a new theoretical analysis on the generalization ability of metric fairness in the margin-based active learning framework. We summarize our results and some closely related prior works in Table 1.

1.3. Overview of our techniques

We sketch the main techniques in this section. From a high level, we leverage the metric-fairness into the celebrated margin-based active learning framework to achieve performance guarantees stated in Theorem 5.

1) Metric-fair learning via convex fairness loss. Given the definition of metric-fairness, it is natural to consider an indicator function as the loss to evaluate whether the fairness constraint is violated for a hypothesis w on a pair $(x, x') \in \mathcal{X} \times \mathcal{X}$:

$$f_\zeta(w; (x, x')) = \begin{cases} 1 & \text{if } \big|w \cdot x - w \cdot x'\big| > \zeta(x, x'), \\ 0 & \text{otherwise.} \end{cases}$$

Yet, such loss function is discrete that is often computationally hard to optimize. Thus, we consider the following surrogate loss that is amenable for optimization:

$$\begin{aligned} &f_\zeta^G(w; (x, x')) \\ &:= \max\big\{0, G(\big|w \cdot x - w \cdot x'\big| - \zeta(x, x')) + 1\big\}, \end{aligned}$$

which is an hinge-loss type upper bound of $f_\zeta(w; (x, x'))$ and very importantly, is convex with respect to w. Now given a set T of instances drawn independently from D_X, it is possible to (arbitrarily) group each two instances to form a set $M(T) \subset \mathcal{X} \times \mathcal{X}$ and examine the empirical fairness loss induced by $M(T)$:

$$\sum_{(x,x') \in M(T)} f_\zeta^G(w; (x, x')).$$

Note that $M(T)$ can be thought of as a set of instance pairs independently drawn from $\mathcal{X} \times \mathcal{X}$, sometimes called a graph matching of T by imaging the instances in T as the nodes of a graph. This would allow us to establish generalization guarantee of metric-fairness, similar to Yona & Rothblum (2018). In our algorithm, we will mainly consider a constraint for the above empirical loss function which can be

evaluated in polynomial time. This is one of the primary components when the algorithm produces a new iterate. By enforcing such convex fairness constraint in all iterations, we obtain the fairness guarantee for all iterates, and hence the final output of the algorithm.

2) Margin-based active learning with fairness: entangling unlabeled and labeled data. The margin-based active learning framework of Balcan et al. (2007) proceeds in an iterative fashion, where in each phase it minimizes an empirical hinge loss to produce a new iterate. The key aspect that differentiates it from passive learning is that in each phase, it draws a bulk of instances from D_X but only queries the labels of those residing a band B, known as localized sampling. In addition, it searches the new iterate in a localized hypothesis space, which is roughly a trust region of the target halfspace w^*. As the algorithm proceeds, such localized hypothesis space shrinks at a geometric rate, hence after a few phases, a good hypothesis can be returned.

We make several key observations. First, the labels are queried only during empirical hinge loss minimization, and the number of labeled instances is such that the empirical hinge loss is a good approximation to the expected loss. Second, the instances are drawn at the beginning of each phase and will be used to construct the surrogate fairness constraint. It turns out that such unlabeled and labeled instances are interleaved during hinge loss minimization, i.e. labeled samples are used to define the hinge loss, while unlabeled ones are used to construct the constraint. Yet, we show that the size of unlabeled samples and the necessary size of labels are almost independent. Therefore, different from prior active learning algorithms, after localized sampling, we also perform a random sampling of the remaining instances and query their labels. This ensures that we only make necessary label queries, and is the key to obtain the poly-logarithmic dependence on $\frac{1}{\alpha}$.

Our treatment on adversarial noise is rather standard. It turns out that the framework inherently is robust to the adversarial noise as long as the noise rate $\eta \leq O(\epsilon)$, as set out in Awasthi et al. (2017). Technically speaking, this comes from the fact that the expected error within the localized sampling region is only constant away from the error rate of w^*, which suffices to establish the desired bound on classification error rate. Moreover, we can show that the additional fairness constraint will not hurt such analysis.

Finally, in order to incorporate the sparsity structure of the underlying halfspace, we consider enforcing an additional ℓ_1-norm constraint in the localized hypothesis space. This narrows down the search space and hence improves the label complexity. At a technical level, the additional ℓ_1-norm constraint significantly reduces the Rademacher complexity (Kakade et al., 2008). Since the ℓ_1-norm constraint may promote a non-sparse halfspace, at the end of each phase, we will perform hard thresholding to ensure sparsity. This will increase the error rate by a constant factor but we can show that overall, it can still be well-controlled, which is a key observation made in Zhang (2018).

1.4. Roadmap

We discuss more related works in Section 2, including fairness and active learning. A concrete problem setup is presented in Section 3. We elaborate on our main algorithms in Section 4, followed by performance guarantees in Section 5. We conclude the paper in Section 6, and defer all proof details to the appendix.

2. Related Works

In this section, we provide a brief review of learning with fairness and active learning.

Fairness. There are two important fairness criteria: statistical fairness and individual fairness. Statistical fairness seeks to stabilize a small number of protected demographic groups (e.g. kids) and then requires some statistical metric be equal across all these groups. The algorithms use a variety of metrics, the most popular of which are the raw positive classification rate (Feldman et al., 2015), the false positive and false negative classification rates (Hardt et al., 2016; Kleinberg et al., 2017), and the positive predictive value (Kleinberg et al., 2017). However, statistical fairness does not provide adequate protection for individuals or a structured subgroup since it examines whether the protected groups receive an average benefit.

In contrast to statistical fairness, individual fairness focuses on specific individuals rather than an average across populations. Dwork et al. (2012) suggested that for each pair of individuals, a metric should be used in such a way that similar individuals should be treated similarly; this is the concept of individual fairness. Based on the notion of individual fairness, a series of algorithms have been developed that integrates the online learning setting and guarantees the individual fairness via an oracle to evaluate fairness violations (Kim et al., 2018; Kearns et al., 2018; Gillen et al., 2018). A very elegant work due to Yona & Rothblum (2018) presented generalization guarantee of fairness under the notion of approximate metric-fairness. For a comprehensive discussion on fairness in machine learning, we refer readers to a recent survey article by Mehrabi et al. (2021).

Active learning. Active learning concerns the scenario where unlabeled data are abundant but labeling could be very expensive. The study of active learning was initiated by Cohn et al. (1994). It, however, turns out that in general, even for the very simple problem of learning halfspaces, active learning may not be able to provide any improvement on label complexity compared to passive learning

(Dasgupta, 2005). In this regard, distributional assumptions on the unlabeled data are often made, such as uniform distribution on a unit ball (Balcan et al., 2007). One line of research considers combining empirical risk minimization with active learning and demonstrates surprising properties beyond label efficiency, such as noise tolerance (Awasthi et al., 2017; Shen, 2021b). Another line of research designs more sophisticated algorithms for improved noise tolerance (Yan & Zhang, 2017; Zhang et al., 2020; Diakonikolas et al., 2020b;a; Shen, 2021a; Zhang & Li, 2021).

Interestingly, Awasthi et al. (2016); Zhang (2018) observed that the margin-based active learning framework inherently is compatible with attribute-efficiency, an important property that has been investigated in learning theory and statistics (Blum, 1990; Blum et al., 1991; Donoho, 2006; Tropp & Wright, 2010; Shen & Li, 2018). The objective of attribute-efficient learning is to learn a sparse model that improves the performance guarantee on sample and label complexity, typically with the hope of obtaining bounds that are logarithmic in the dimension. More recently, Shen & Zhang (2021) developed an attribute-efficient active learning algorithm that is robust to the malicious noise where both instances and labels can be adversarially corrupted (Valiant, 1985).

3. Preliminaries

For a vector $w \in \mathbb{R}^d$, we denote its ℓ_2 and ℓ_1-norm by $\|w\|_2$ and $\|w\|_1$, respectively. We use $\|w\|_0$ to count the number of non-zero elements of w. We will denote by $\mathcal{B}_2(w, r)$ the ℓ_2-ball centering at w with radius r, and by $\mathcal{B}_1(w, \rho)$ the ℓ_1-ball centering at w with radius ρ. We define the angle between two vectors w_1, w_2 in $\mathbb{R}^d$ as $\theta(w_1, w_2) = \arccos\left(\frac{w_1 \cdot w_2}{\|w_1\|_2 \|w_2\|_2}\right)$.

The letter c and its subscript variants such as c_1, c_2, c_3 are reserved for specific constants. The letters κ and $\bar{c}$ are also reserved constants. Readers may refer to Appendix A for the detailed values. The capital letter K and its subscript variants are also constants, but their values may change from appearance to appearance.

Recall that we denote the instance space by $\mathcal{X} \subset \mathbb{R}^d$, the label space by $\mathcal{Y}$ and the class of t-sparse homogeneous halfspaces by $\mathcal{H} := \{x \mapsto \text{sign}(w \cdot x) : w \in \mathbb{R}^d, \|w\|_2 = 1, \|w\|_0 \leq t\}$. Note that a halfspace is non-sparse if $t = d$. Hence, this is a more general definition of the underlying hypothesis class.

In our algorithm, we will often choose unlabeled instances from D_X conditional on a sampling region $B := \{x \in \mathbb{R}^d : |w \cdot x| \leq b\}$. This can be done by repeatedly calling EX_D until seeing an instance x lying in B; this process is referred to as rejection sampling. The distribution of D_X conditional on the event $x \in B$ is denoted by $D_{X|B}$.

We will consider a scaled hinge loss as a proxy to the $0/1$ classification error,

$$\ell_\tau\big(w; (x, y)\big) := \max\Big\{0, 1 - \frac{yw \cdot x}{\tau}\Big\}. \tag{1}$$

Note that even with the scaling parameter $\tau > 0$, such hinge loss still upper bounds the $0/1$ classification error.

The expected scaled hinge loss for some distribution $\bar{D}$ over $\mathcal{X} \times \mathcal{Y}$ is thus given by

$$\ell_\tau(w; \bar{D}) := \mathbb{E}_{(x,y)\sim\bar{D}}[\ell_\tau(w; (x, y))]. \tag{2}$$

Further, we will consider the empirical scaled hinge loss with respect to a sample set $S \subset \mathcal{X} \times \mathcal{Y}$,

$$\ell_\tau(w; S) := \frac{1}{|S|} \sum_{(x,y)\in S} \ell_\tau(w; (x, y)). \tag{3}$$

In the sequel, we summarize useful definitions for fairness.

Definition 6 (Metric-fairness loss)**.** The metric-fairness loss on a pair (x, x') with respect to a metric $\zeta : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is as follows:

$$f_\zeta(w; (x, x')) := \mathbf{1}\Big\{\big|w \cdot x - w \cdot x'\big| > \zeta(x, x') + \gamma\Big\}, \tag{4}$$

where $\mathbf{1}\{E\}$ is the indicator function which outputs 1 if the event E holds and 0 otherwise.

The expected metric-fairness loss with respect to the distribution D_X is thus given by

$$f_\zeta(w; D_X) := \mathbb{E}_{(x,x')\sim D_X \times D_X}\big[f_\zeta(w; (x, x'))\big]. \tag{5}$$

We will establish uniform convergence of metric-fairness loss through Rademacher complexity. One key requirement is that the empirical pairs (x, x') need to be independent draws according to $D_X \times D_X$. However, in our algorithmic design, we will not do so. Rather, we draw a set T of instances and construct the set of pairs through the notion of matching in graph theory by thinking of T as a fully connected graph with instances being the vertices, an elegant idea due to Yona & Rothblum (2018).

Definition 7 (Matching)**.** Given a set T of instances in $\mathcal{X}$, a matching $M(T)$ is a set of instance pairs without common instances. In addition, all instances in T are contained in one pair in $M(T)$.

Equipped with a matching $M(T)$, it is possible to evaluate the degree of violation of the fairness constraint for a given hypothesis w on a given set T. For example, we may consider the empirical metric-fairness loss $f_\zeta(w; M(T)) := \frac{1}{|M(T)|} \sum_{(x,x')\in M(T)} f_\zeta(w; (x, x'))$. However, $f_\zeta(w; (x, x'))$ is a nonconvex function that is intractable to optimize. Yona & Rothblum (2018)

considered a function of the form $\hat{f}_\zeta(w;(x,x')) := \max\{0, |w\cdot x - w\cdot x'| - \zeta(x,x')\}$ which is convex but not always an upper bound of the $0/1$ metric-fairness loss (4). This resulted in technical complications in their uniform convergence.

We propose to use the following surrogate function.

Definition 8 (Surrogate metric-fairness loss)**.** The surrogate metric-fairness loss on a pair (x,x') with respect to a metric $\zeta : \mathcal{X}\times\mathcal{X}\to[0,1]$ is as follows:

$$f^G_\zeta(w;(x,x')) := \max\{0, G(|w\cdot x - w\cdot x'| - \zeta(x,x')) + 1\}, \quad (6)$$

where $G > 0$ is a parameter.

It is easy to see that the surrogate loss is convex with respect to w and is always an upper bound of the $0/1$ loss in Definition 6. Thus, given an instance set T, we can build an arbitrary matching $M(T)$ and consider the convex fairness loss over $M(T)$ as follows:

$$f^G_\zeta(w;M(T)) := \frac{1}{|M(T)|}\sum_{(x,x')\in M(T)} f^G_\zeta(w;(x,x')). \quad (7)$$

Since our algorithm and analysis hold for any matching of T, we will often write $f^G_\zeta(w;T)$ in place of $f^G_\zeta(w;M(T))$.

4. Main Algorithms

We describe the main algorithms in this section. For the purpose of exposition, we will start with a basic algorithm that works under the noise-free data and without sparsity pattern of the underlying halfspace. We then present an attribute-efficient algorithm when the hypothesis class is t-sparse halfspaces. For both algorithms, there is no particular treatment on the adversarial noise; it only enters our theoretical analysis.

4.1. Hyper-parameter setting

Our algorithms proceed in phases. In each phase k, we set

$$r_k = 2^{-k-3},\ b_k = \bar{c}\cdot r_k,\ \rho_k = \sqrt{2t}\cdot r_k,\ \tau_k = \kappa\cdot b_k, \quad (8)$$

where ρ_k is used only in Algorithm 2. The failure probability $\delta_k := \frac{\delta}{(k+1)(k+2)}$. We will draw a set of instances T_k by calling EX_D for n_k times, where

$$n_k = K_1\big(\frac{1}{\alpha^2} + \frac{1}{b_k}\big)t\log^4\frac{d}{\epsilon\delta_k}. \quad (9)$$

We will then query the label of some instances in T_k by calling EX^Y_D for m_k times, where

$$m_k = K_2\cdot t\log^3\frac{d}{\alpha\delta_k}\cdot\log d. \quad (10)$$

Algorithm 1 Active Learning of Halfspaces with Approximate Metric-Fairness

Require: Target classification error rate ϵ, failure probability δ, fairness metric function $\zeta(\cdot,\cdot)$, fairness error rate α, sample generation oracle EX_D, label revealing oracle EX^Y_D.

Ensure: A halfspace $\hat{w}$ with $\mathrm{err}_D(\hat{w}) \le \epsilon$ and is also α-approximately metric-fair.

1: $k_{\max} \leftarrow \log(\frac{\pi}{128c_1\epsilon})$.
2: **for** $k = 1, 2, \ldots, k_{\max}$ **do**
3: $T_k \leftarrow$ independently draw n_k instances from D_X.
4: Build a matching $M(T_k)$ and construct W_k.
5: $S_k \leftarrow$ randomly sample m_k instances in $T_k \cap B_k$ and query their labels.
6: Find $w_k \in W_k$ such that

$$\ell_{\tau_k}(w_k;S_k) \le \arg\min_{w\in W_k}\ell_{\tau_k}(w;S_k) + \kappa.$$

7: **end for**
return $\hat{w} \leftarrow w_{k_{\max}}$.

Recall $\kappa, \bar{c} > 0$ are reserved constants (see Appendix A), and $K_1, K_2 > 0$ are constants that we will not particularly track their values. It is also worth noting that n_k and m_k are referred to as the sample size and label size at phase k.

4.2. Active learning of general halfspaces with fairness

Our algorithm proceeds in multiple phases. Fix a phase $k \ge 1$. The basic metric-fair active learning, Algorithm 1, consists of two major stages: sampling from a region B_k, called localized sampling, and minimizing a hinge loss under certain constraint set W_k. Since the sample generation oracle only returns instances drawn from D_X, we need to call it sufficient times to obtain an instance set T_k, and identify those in B_k.

The localized sampling region B_k is defined as follows:

$$B_k := \begin{cases}\mathbb{R}^d & k = 1,\\ \{x : |w_{k-1}\cdot x| \le b_k\} & k \ge 2.\end{cases} \quad (11)$$

Intuitively, as the algorithm proceeds, the iterate w_{k-1} will be very close to the target halfspace w^*, and thus only instances close to the decision boundary of w_{k-1} are informative in the sense that w_{k-1} may disagree with w^* on their labels – this is why we only query the labels in the band B_k. Note that such setting is standard in margin-based active learning (Awasthi et al., 2017).

The design of W_k is more delicate, as we need to ensure that the iterates are metric-fair. Without the fairness consideration, it is known in the active learning literature that we

can choose W_k same as $\mathcal{Q}_k$, where

$$\mathcal{Q}_k = \begin{cases} \mathcal{B}_2(0,1) & k=1, \\ \mathcal{B}_2(0,1) \cap \mathcal{B}_2(w_{k-1}, r_k) & k \geq 2. \end{cases} \tag{12}$$

The constraint set $\mathcal{Q}_k$ has two properties with a high probability: 1) it is shrinking; and 2) the target halfspace w^* keeps lying in W_k for all $k \geq 1$ (see Proposition 21).

To account for the fairness property, we need additional constraint. We will consider

$$\mathcal{M}_k = \left\{ w \in \mathbb{R}^d : f_\zeta^{G_k}(w; T_k) \leq \frac{\alpha}{2} \right\}. \tag{13}$$

Recall that $f_\zeta^{G_k}(w; T_k)$ is a convex function that upper bounds the $0/1$ fairness error. This will be a useful fact for us to show generalization ability of the metric-fairness property. The parameter $G_k > 0$ can be adaptively selected. Yet, we find that $G_k = 1$ works in our analysis. It is unclear whether a more involved choice would improve the performance guarantees; we leave it as our future study. We would also like to mention that Yona & Rothblum (2018) proposed to use a surrogate loss $\hat{f}_\zeta(w; T_k) := \frac{1}{|M(T_k)|} \sum_{M(T_k)} \max\{0, |w \cdot x - w \cdot x'| - \zeta(x, x')\}$. This is convex as well but the trouble is that it is not always an upper bound of the $0/1$ fairness error defined in Definition 6. Thus, there are technical complications in the analysis of Yona & Rothblum (2018).

Consequently, we will consider a constraint set for the candidate halfspaces as follows:

$$W_k = \mathcal{M}_k \cap \mathcal{Q}_k. \tag{14}$$

Note that any $w \in W_k$ is $\frac{\alpha}{2}$-approximately metric-fair on the instance set T_k.

Then we need to find a good halfspace w_k. To this end, we minimize the hinge loss $\ell_{\tau_k}(w; S_k)$. The labeled instance set S_k is chosen in a very involved manner. Given T_k drawn at the beginning, we first identify those residing B_k, denoted by $T_k' := T_k \cap B_k$. While prior works query the labels of all instances in T_k' (Balcan et al., 2007; Zhang, 2018), we will first perform a random sampling from T_k' to form an instance set of size m_k, and query EX_D^Y on their labels. As will be clear in the analysis, m_k is orders of magnitude smaller than the size of T_k' – a key step to ensure the announced label complexity. In fact, had we not performed the random sampling, our label complexity would be proportional to $\frac{1}{\alpha^2}$, which is exponentially more than what we obtained. Technically speaking, the reason that we only need a small amount of labeled data is that m_k labeled instances suffice to guarantee uniform convergence to the expected hinge loss, which will imply a small classification error. On the other spectrum, we do need a large amount of unlabeled data (i.e. T_k) to guarantee uniform convergence to the expected fairness error. Therefore, the algorithm entangles the unlabeled and labeled data yet disentangles their sizes. Such idea of random sampling also appeared in a very recent work of Shen & Zhang (2021) but was motivated in a quite different context: in that work, they need to draw a bulk of instances T_k to detect malicious instances (Valiant, 1985), and then sample a small amount for labeling to obtain near-optimal label complexity.

Equipped with the labeled instance set S_k and the convex constraint set W_k, we optimize the empirical hinge loss up to a constant $\kappa > 0$, which is computationally efficient. Here, the hinge loss is parameterized by a Lipschitz coefficient τ_k that shrinks exponentially with respect to the iteration number k; this will be useful to control the label complexity. A potential trouble is that whether such convex program is feasible. We give an affirmative answer; in fact, we show that the target halfspace is a feasible solution, namely, $w^* \in W_k$ for all phases k. Observe that this immediately implies that after $O(\log \frac{1}{\epsilon})$ iterations, we have $\|w_k - w^*\|_2 \leq \epsilon$ due to the shrinking ℓ_2-norm constraint we imposed in $\mathcal{Q}_k$ and the setting $r_k = \Theta(2^{-k})$.

Notably, a natural characteristic of our algorithm is its noise tolerance. When the label is corrupted by adversarial noise during the training phase, the shrinking sampling region B_k well-controls the amplitude of the noise. Therefore, Algorithm 1 simultaneously guarantees the metric fairness and tolerance to the adversarial noise .

4.3. Active learning of sparse halfspaces with fairness

Now we present Algorithm 2, which leverages the prior knowledge that the underlying hypothesis class is t-sparse halfspaces to achieve attribute efficiency.

Inspired by Zhang (2018), we modify the constraint set $\mathcal{Q}_k$ in (12) as follows to incorporate the sparsity structure:

$$\mathcal{Q}_k = \begin{cases} \mathcal{B}_2(0,1) \cap \mathcal{B}_1(0, \sqrt{t}) & k=1, \\ \mathcal{Q}_1 \cap \mathcal{B}_2(w_{k-1}, r_k) \cap \mathcal{B}_1(w_{k-1}, \rho_k) & k \geq 2. \end{cases} \tag{15}$$

The key difference from the $\mathcal{Q}_k$ that we used in Algorithm 1 is that the constraint set $\mathcal{Q}_k$ belongs to the intersection between ℓ_1-norm ball and ℓ_2-norm balls centering at w_{k-1}, which will be useful to establish label complexity that is logarithmic in the dimension. Furthermore, we also narrow down the constraint set $\mathcal{Q}_k$ such that it is a subset of $\mathcal{Q}_1$ for all $k \geq 1$. This will simplify our analysis for the generalization of metric-fairness (but not vital).

It is worth mentioning that the solution of hinge loss minimization v_k is not always t-sparse. For technical reasons, we will perform a hard thresholding step $\mathrm{P}_t(v_k)$ which sets all but the t largest (in magnitude) elements of v_k to zero, followed by an ℓ_2-normalization. We note that though hard

Algorithm 2 Active Learning of Sparse Halfspaces with Approximate Metric-Fairness

Require: Sparsity parameter t, target classification error rate ϵ, failure probability δ, fairness metric function $\zeta(\cdot,\cdot)$, fairness error rate α, sample generation oracle EX_D, label revealing oracle EX_D^Y.

Ensure: A halfspace $\hat{w}$ with $\mathrm{err}_D(\hat{w}) \le \epsilon$ and is also α-approximately metric-fair.

1: $k_{\max} \leftarrow \log(\frac{\pi}{128 c_1 \epsilon})$.
2: **for** $k = 1, 2, \ldots, k_{\max}$ **do**
3: $T_k \leftarrow$ independently draw n_k instances from D_X.
4: Build a matching $M(T_k)$ and construct W_k.
5: $S_k \leftarrow$ randomly sample m_k instances in $T_k \cap B_k$ and query their labels.
6: Find $v_k \in W_k$ such that

$$\ell_{\tau_k}(v_k; S_k) \le \arg\min_{w \in W_k} \ell_{\tau_k}(w; S_k) + \kappa.$$

7: $w_k \leftarrow \frac{\mathrm{P}_t(v_k)}{\|\mathrm{P}_t(v_k)\|_2}$.
8: **end for**
return $\hat{w} \leftarrow v_{k_{\max}}$.

thresholding is a non-convex projection, it is still possible to control the resultant deviation; see Proposition 21.

Another notable aspect of Algorithm 2 is that the returned hypothesis is $v_{k_{\max}}$ rather than $w_{k_{\max}}$. Though it is possible to show that $w_{k_{\max}}$ enjoys PAC guarantee as well as $v_{k_{\max}}$ does, we find it technically difficult to prove its fairness guarantee, due to the non-convexity nature of the hard thresholding operation. Therefore, the returned halfspace is *not* t-sparse, yet with PACF guarantee and attribute-efficiency.

5. Performance Guarantees

We provide the performance guarantee for our algorithms in this section. Since Algorithm 1 is a special case of Algorithm 2, we state the main results of the latter first, and the guarantees of Algorithm 1 follow as an immediate corollary.

Theorem 9 (Main result). *Suppose Assumptions 1, 2, 3 are satisfied, and w^* is such that $f_\zeta(w^*; D_X) = 0$. For any given $\epsilon, \delta, \alpha \in (0,1)$, if $\eta \le c_0 \epsilon$ for sufficiently small constant $c_0 > 0$, then the following holds. With probability $1-\delta$, Algorithm 2 runs in polynomial time and returns a halfspace $\hat{w}$ such that $\mathrm{err}_D(\hat{w}) \le \epsilon$. In addition, $\hat{w}$ is α-approximately metric-fair. The sample complexity is $O\big(\big(\frac{1}{\alpha^2} + \frac{1}{\epsilon}\big) \cdot t \log^4 \frac{d}{\epsilon\delta} \cdot \log \frac{1}{\epsilon}\big)$, and the label complexity is $O\big(t \log^3 \frac{d}{\alpha\delta} \cdot \log d \cdot \log \frac{1}{\epsilon}\big)$.*

Observe that the fairness metric $\zeta(\cdot,\cdot)$ enters our analysis through its maximal value, which is assumed to be 1 in Definition 1. It is straightforward to reproduce our analysis for any universally bounded metric function; we leave it to interested readers.

Corollary 10. *Suppose Assumptions 1 and 2 are satisfied, and w^* is such that $f_\zeta(w^*; D_X) = 0$. For any given $\epsilon, \delta, \alpha \in (0,1)$, if $\eta \le c_0 \epsilon$ for sufficiently small constant $c_0 > 0$, then the following holds. With probability $1-\delta$, Algorithm 1 runs in polynomial time and returns a halfspace $\hat{w}$ such that $\mathrm{err}_D(\hat{w}) \le \epsilon$. In addition, $\hat{w}$ is α-approximately metric-fair. The sample complexity is $O\big(\big(\frac{1}{\alpha^2} + \frac{1}{\epsilon}\big) \cdot d \log^4 \frac{d}{\epsilon\delta} \cdot \log \frac{1}{\epsilon}\big)$, and the label complexity is $O\big(d \log^3 \frac{d}{\alpha\delta} \cdot \log d \cdot \log \frac{1}{\epsilon}\big)$.*

5.1. Proof sketch of Theorem 9

To prove our main result, we show the generalization bound of metric fair loss and then establish the PAC guarantee.

To begin with, we demonstrate a generalization bound of our metric fair loss $f_\zeta^{G_k}(w; (x, x'))$.

Lemma 11 (Lemma 14, informal). *Consider phase k of Algorithm 2. Let $Z_k := \max_{(x,x') \in M(T_k)} \zeta(x, x')$, $X_k := \max_{x \in T_k} \|x\|_\infty$, $\Pi_k := G_k(2\sqrt{t} X_k + Z_k)$. If $|T_k| \ge K \cdot \frac{1}{(\alpha')^2}(G_k^2 X_k^2 t \log d + G_k^2 Z_k^2 + \Pi_k^2) \log \frac{1}{\delta'}$ for some constant $K > 0$, then with probability $1-\delta'$ over the draw of T_k, the following holds for any $\alpha' > 0$:*

$$\sup_{w: \|w\|_1 \le \sqrt{t}} \left| f_\zeta^{G_k}(w; D_X) - f_\zeta^{G_k}(w; T_k) \right| \le \alpha'.$$

Such maximal difference between empirical and expected loss functions follows from uniform convergence via Rademacher complexity (Bartlett & Mendelson, 2001). The primary challenge is to estimate the Rademacher complexity for the underlying hypothesis class. We show that it is possible to upper bound the maximal value of the function $f_\zeta^{G_k}(w; (x, x'))$ in view of the constraint set W_k and our distributional assumption on D_X. This in conjunction with the fact that it is G_k-Lipschitz implies the desired result.

By setting $\alpha' = \frac{\alpha}{2}$ in Lemma 11 and combining with the constraint $f_\zeta^{G_k}(w; T_k) \le \frac{\alpha}{2}$ and the observation that $f_\zeta^{G_k}(w; (x, x'))$ is always an upper bound of the $0/1$ fairness error $f_\zeta(w; (x, x'))$, we obtain the approximate metric-fairness guarantee; see Appendix B for the full proof.

Theorem 12 (Generalization of fairness). *Consider phase k of Algorithm 2. With probability $1-\delta_k$ over the draw of T_k, all $w \in W_k$ are α-approximately metric-fair with respect to $\zeta(\cdot,\cdot)$ provided that $|T_k| \ge K \cdot \frac{1}{\alpha^2}(t \log^4 \frac{d}{\delta_k})$ for some constant $K > 0$. In particular, the hinge loss minimizer v_k is α-approximately metric-fair.*

We now discuss the PAC guarantee stated in Theorem 9. Let $T_k' = T_k \cap B_k$ and we imagine that T_k' is labeled by EX_D^Y. In this way, the samples in T_k' are independent draws

from D_k, the distribution D conditional on the event $x \in B_k$. Consequently, S_k is a subset that is randomly sampled from T'_k. We show that with the setting of n_k and m_k, the following inequalities hold:

$$\sup_{w \in \mathcal{Q}_k} \left| \ell_{\tau_k}(w; S_k) - \ell_{\tau_k}(w; T'_k) \right| \leq \kappa,$$
$$\sup_{w \in \mathcal{Q}_k} \left| \ell_{\tau_k}(w; T'_k) - \ell_{\tau_k}(w; D_k) \right| \leq \kappa,$$
$$\sup_{w \in \mathcal{Q}_k} \left| \ell_{\tau_k}(w; D_k) - L_{\tau_k}(w; D_{X|B_k}) \right| \leq \kappa,$$

where $L_{\tau_k}(w; D_{X|B_k}) := \mathbb{E}_{x \sim D_{X|B_k}}[\ell_{\tau_k}(w; (x, \text{sign}(w^* \cdot x)))]$. The first inequality shows that the random sampling of S_k from T'_k preserves the hinge loss, which is crucial for obtaining improved label complexity since we only need to annotate S_k whose size is m_k that is much less that $|T'_k|$ (which is roughly $\Theta(b_k n_k)$). The second inequality is not surprising due to uniform convergence. The last inequality is very useful to handle the adversarial noise, since it asserts that the hinge loss on the corrupted distribution D_k does not deviate far from that on the clean distribution. Combining them, we can show that $L_{\tau_k}(w; D_{X|B_k}) \approx \ell_{\tau_k}(w; S_k)$ up to a constant additive factor. This suffices to establish the PAC guarantee in view of standard results from margin-based active learning; see Appendix C.

Combining the PAC guarantee and Theorem 12, we obtain the PACF guarantee. Finally, recall that the sample complexity refers to the total number of calls to EX_D, which is the sum of all $n_k := |T_k|$; the label complexity refers to the total number of calls to EX_D^Y, which is the sum of all $m_k := |S_k|$. This completes the proof of Theorem 9 by observing our setting on n_k and m_k in Section 4.1.

6. Conclusion and Future Works

In this paper, we presented the first computationally efficient active learning algorithm with the property of approximate metric-fairness. The core idea is to interleave unlabeled and labeled data in a delicate way to obtain exponential improvement on the label complexity while retaining metric-fairness for the potential hypotheses. Our analysis is based on the presumption that a perfectly metric-fair hypothesis exists in the given hypothesis class. It would be useful to consider a weaker condition that such target hypothesis is only approximately metric-fair. It is also important to investigate whether we can improve the sample complexity in terms of the dependence on the fairness error rate, say proportional to $\frac{1}{\alpha}$.

Acknowledgements

We thank the anonymous reviewers and meta-reviewer for valuable comments on improving the notation and proof structure. This work is supported by NSF-IIS-1948133 and the startup funding from Stevens Institute of Technology.

References

Awasthi, P., Balcan, M., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 152–192, 2016.

Awasthi, P., Balcan, M., and Long, P. M. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50:1–50:27, 2017.

Balcan, M. and Long, P. M. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 288–316, 2013.

Balcan, M., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 65–72, 2006.

Balcan, M., Broder, A. Z., and Zhang, T. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pp. 35–50, 2007.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pp. 224–240, 2001.

Blum, A. Learning boolean functions in an infinite atribute space. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pp. 64–72, 1990.

Blum, A., Hellerstein, L., and Littlestone, N. Learning in the presence of finitely or infinitely many irrelevant attributes. In *Proceedings of the 4th Annual Workshop on Computational Learning Theory*, pp. 157–166, 1991.

Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

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

Dasgupta, S. Coarse sample complexity bounds for active learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 235–242, 2005.

Dasgupta, S. The two faces of active learning. In *Proceedings of the 20th International Conference On Algorithmic Learning Theory*, pp. 1, 2009.

Dasgupta, S., Kalai, A. T., and Monteleoni, C. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Learning Theory*, pp. 249–263, 2005.

Diakonikolas, I., Kane, D. M., Kontonis, V., Tzamos, C., and Zarifis, N. A polynomial time algorithm for learning halfspaces with Tsybakov noise. *CoRR*, abs/2010.01705, 2020a.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with Massart noise under structured distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pp. 1486–1513, 2020b.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 6478–6490, 2021.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Dwork, C. and Ilvento, C. Fairness under composition. In *Proceedings of the 10th Innovations in Theoretical Computer Science Conference*, pp. 33:1–33:20, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science*, pp. 214–226, 2012.

Dwork, C., Ilvento, C., and Jagadeesan, M. Individual fairness in pipelines. In *Proceedings of the 1st Symposium on Foundations of Responsible Computing*, pp. 7:1–7:22, 2020a.

Dwork, C., Ilvento, C., Rothblum, G. N., and Sur, P. Abstracting fairness: Oracles, metrics, and interpretability. In *Proceedings of the 1st Symposium on Foundations of Responsible Computing*, pp. 8:1–8:16, 2020b.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21st International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Gillen, S., Jung, C., Kearns, M. J., and Roth, A. Online learning with an unknown fairness metric. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 2605–2614, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pp. 3315–3323, 2016.

Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 793–800, 2008.

Kearns, M. J., Schapire, R. E., and Sellie, L. Toward efficient agnostic learning. In *Proceedings of the 5th Annual Conference on Computational Learning Theory*, pp. 341–352, 1992.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2569–2577, 2018.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness through computationally-bounded awareness. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 4847–4857, 2018.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, pp. 43:1–43:23, 2017.

Littlestone, N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm (extended abstract). In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pp. 68–77, 1987.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3156–3164, 2018.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4051–4060, 2019.

Lovász, L. and Vempala, S. S. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007a.

Lovász, L. and Vempala, S. S. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007b.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):115:1–115:35, 2021.

Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Shen, J. On the power of localized Perceptron for label-optimal learning of halfspaces with adversarial noise. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9503–9514, 2021a.

Shen, J. Sample-optimal PAC learning of halfspaces with malicious noise. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9515–9524, 2021b.

Shen, J. and Li, P. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

Shen, J. and Zhang, C. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1072–1113, 2021.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tropp, J. A. and Wright, S. J. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

Valiant, L. G. A theory of the learnable. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, pp. 436–445, 1984.

Valiant, L. G. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pp. 560–566, 1985.

Vempala, S. S. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM*, 57(6):32:1–32:14, 2010.

Yan, S. and Zhang, C. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 1056–1066, 2017.

Yona, G. and Rothblum, G. N. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5666–5674, 2018.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333, 2013.

Zhang, C. Efficient active learning of sparse halfspaces. In *Proceedings of the 31st Annual Conference on Learning Theory*, pp. 1856–1880, 2018.

Zhang, C. and Li, Y. Improved algorithms for efficient active learning halfspaces with Massart and Tsybakov noise. In *Proceedings of the 34th Annual Conference on Learning Theory*, pp. 4526–4527, 2021.

Zhang, C., Shen, J., and Awasthi, P. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 7184–7197, 2020.

A. Summary of Useful Notations and Reserved Parameters

We recall the following fairness loss functions on a pair $(x, x') \in \mathcal{X} \times \mathcal{X}$:

$$f_\zeta(w; (x, x')) = \mathbf{1}\Big\{\big|w \cdot x - w \cdot x'\big| - \zeta(x, x') > 0\Big\},$$
$$f_\zeta^G(w; (x, x')) = \max\big\{0, G(\big|w \cdot x - w \cdot x'\big| - \zeta(x, x')) + 1\big\}.$$

Note that the latter is always an upper bound of the former, and that the latter is convex with respect to w.

Given a set T of instances drawn from the distribution D_X and its matching $M(T)$, we defined

$$f_\zeta(w; M(T)) = \frac{1}{\big|M(T)\big|} \sum_{(x,x') \in M(T)} f_\zeta(w; (x, x')), \qquad f_\zeta(w; D_X) = \mathbb{E}_{(x,x') \sim D_X \times D_X}\big[f_\zeta(w; (x, x'))\big].$$

Note that since $M(T)$ is a matching, we have $\big|M(T)\big| = \frac{1}{2}|T|$. Since an arbitrary matching works in our analysis, we often omit the specification and write

$$f_\zeta(w; T) := f_\zeta(w; M(T)).$$

Likewise, we can define $f_\zeta^G(w; T)$ and $f_\zeta^G(w; D_X)$.

We recall the following scaled hinge loss functions on a pair (x, y) in $\mathcal{X} \times \mathcal{Y}$:

$$\ell_\tau\big(w; (x, y)\big) = \max\Big\{0, 1 - \frac{yw \cdot x}{\tau}\Big\}.$$

Given a distribution $\bar{D}$ on $\mathcal{X} \times \mathcal{Y}$, let

$$\ell_\tau(w; \bar{D}) := \mathbb{E}_{(x,y) \sim \bar{D}}[\ell_\tau(w; (x, y))].$$

Further, we will consider the hinge loss with respect to a set S of labeled instances,

$$\ell_\tau(w; S) := \frac{1}{|S|} \sum_{(x,y) \in S} \ell_\tau(w; (x, y)).$$

For a given phase $k \geq 1$, we collect useful notations in Table 2.

Table 2. Notations used in phase k of Algorithm 2.

B_k	the band $\{x : \lvert w_{k-1} \cdot x\rvert \leq b_k\}$
$D_{X\lvert B_k}$	the marginal distribution of D_X conditioned on the even $x \in B_k$
D_k	the joint distribution D conditioned on the even $x \in B_k$
T_k	a set of instances drawn from D_X
T_k'	$T_k \cap B_k$ (to appear in our analysis)
S_k	a subset of T_k' and is labeled by EX_D^Y
$M(T_k)$	a matching of T_k by viewing it as a graph
$\mathcal{Q}_k$	$\{w : \lVert w\rVert_2 \leq 1, \lVert w\rVert_1 \leq \sqrt{t}, \lVert w - w_{k-1}\rVert_2 \leq r_k, \lVert w - w_{k-1}\rVert_1 \leq \rho_k\}$
$\mathcal{M}_k$	$\{w : f_\zeta^G(w; T_k) \leq \frac{\alpha}{2}\}$
W_k	$\mathcal{Q}_k \cap \mathcal{M}_k$
X_k	$\max_{x \in T_k} \lVert x\rVert_\infty$

Constants. We clarify the choices of absolute constants. The constants $c_1, c_2, c_3, c_4, c_5, c_6$ are specified in Lemma 23, Lemma 24, and Lemma 25. The constant $\kappa \leq \frac{\pi}{2^{12} c_2}$ and $\bar{c}$ are jointly chosen. Let $g(a) := \big[\kappa c_6 a + \frac{c_3 \pi}{4} \exp\big(-\frac{c_4 a}{4\pi}\big)\big] c_2$. We set $\kappa = \exp(-a)$ and plug into $g(a)$. It then is easy to see that $g(a)$ is continuous and tends to zero as a goes to infinity. Thus, there must exist $\bar{c}$ such that $g(\bar{c}) \leq \frac{\pi}{2^8}$; in addition, $\bar{c}$ is an absolute constant since all coefficients in $g(a)$ are constants. We then let $\kappa = \min\{\exp(-\bar{c}), \frac{\pi}{2^{12} c_2}\}$.

B. Approximate Metric-Fairness Guarantee

Recall that $W_k = \mathcal{M}_k \cap \mathcal{Q}_k$. We will first show that the target halfspace $w^* \in \mathcal{M}_k$. Hence, the hinge loss minimization problem is always feasible. We then show that as long as we draw sufficient number of instances, any feasible solution is approximate metric-fair (Theorem 12).

Lemma 13. *Consider any phase k of Algorithm 2. The target halfspace w^* is contained $\mathcal{M}_k$.*

Proof. Since we consider the realizable setting, we have $f_\zeta(w^*; D_X) = 0$. Namely, $\left|w \cdot x - w \cdot x'\right| \leq \zeta(x, x')$ holds almost surely. This implies that $f_\zeta^{G_k}(w^*; T_k) = 0 \leq \frac{\alpha}{2}$ almost surely. □

The following theorem states that with our constructed constraint set W_k, any $w \in W_k$ is α-approximate metric fair.

B.1. Proof of Theorem 12

Proof. We will mainly use the result in Lemma 14. By our setting, we have $G_k = 1$, $Z_k = 1$, $X_k = \Theta(\log \frac{dn_k}{\delta_k})$. Therefore, when $n_k \geq \tilde{\Omega}\big(\frac{1}{\alpha^2}(t \log^3 d) \log \frac{1}{\delta_k}\big)$, we have

$$\sup_{w \in W_k} \left| f_\zeta^{G_k}(w; D_X) - f_\zeta^{G_k}(w; T_k) \right| \leq \frac{\alpha}{2}. \tag{16}$$

We now upper bound the expected metric-fairness loss when instances are drawn from D_X in phase k. Note that for all $w \in W_k$, we have

$$\begin{aligned} f_\zeta(w; D_X) &= \mathbb{E}_{(x,x') \sim D_X \times D_X} \left[f_\zeta(w; (x, x')) \right] \\ &\leq \mathbb{E}_{(x,x') \sim D_X \times D_X} \left[f_\zeta^{G_k}(w; (x, x')) \right] \\ &\leq f_\zeta^{G_k}(w; T_k) + \frac{\alpha}{2} \\ &\leq \alpha, \end{aligned}$$

where the first inequality follows from our construction of $f_\zeta^G(w; (x, x'))$ which always upper bounds $f_\zeta(w; (x, x'))$, the second inequality follows from (16), and the last inequality follows from the construction of the constraint W_k which ensures $f_\zeta^{G_k}(w; T_k) \leq \frac{\alpha}{2}$. □

B.2. Uniform convergence of metric-fairness loss

Lemma 14 (Formal statement of Lemma 11)**.** *Consider phase k of Algorithm 2. Let $Z_k := \max_{(x,x') \in M(T_k)} \zeta(x, x')$ and let n be the size of T_k. With probability $1 - \delta'$ over the draw of T_k, the following holds:*

$$\sup_{w : \|w\|_1 \leq \sqrt{t}} \left| f_\zeta^{G_k}(w; D_X) - f_\zeta^{G_k}(w; T_k) \right| \leq 4 G_k X_k \sqrt{\frac{2t \log(2d)}{n}} + 2 G_k \sqrt{\frac{Z_k^2}{n}} + \Pi_k \sqrt{\frac{8 \log(2/\delta')}{n}},$$

where $\Pi_k = G_k(2\sqrt{t} X_k + Z_k) + 1$. Therefore, for any $\alpha' > 0$,

$$\sup_{w : \|w\|_1 \leq \sqrt{t}} \left| f_\zeta^{G_k}(w; D_X) - f_\zeta^{G_k}(w; T_k) \right| \leq \alpha'$$

as soon as $n \geq K \cdot \frac{1}{(\alpha')^2}(G_k^2 X_k^2 t \log d + G_k^2 Z_k^2 + \Pi_k^2) \log \frac{1}{\delta'}$ for some constant $K > 0$.

Proof. We consider the following hypothesis class:

$$\mathcal{G} := \left\{ (x, x') \mapsto \left| w \cdot x - w \cdot x' \right| - \zeta(x, x') : w \in W_k \right\}.$$

Observe that $f_\zeta^{G_k} = q \circ g$, where $g \in \mathcal{G}$ and $q(a) = \max\{0, G_k \cdot a + 1\}$. Let $\mathcal{F}$ be the hypothesis class of $q \circ g$. Since $q(\cdot)$ is a G_k-Lipschitz function, it is known that

$$\mathcal{R}_n(\mathcal{F}) \leq G_k \cdot \mathcal{R}_n(\mathcal{G}),$$

where $\mathcal{R}_n(\cdot)$ denotes the empirical Rademacher complexity of a sample set with size n. Let $\mathcal{L} := \{x \mapsto w \cdot x : \|w\|_1 \leq \sqrt{t}\}$. By Claim 2.17 of Yona & Rothblum (2018), we have

$$\mathcal{R}_n(\mathcal{G}) \leq 4\mathcal{R}_n(\mathcal{L}) + 2\sqrt{\frac{\max_{(x,x')\in M(T_k)} \zeta^2(x,x')}{n}}.$$

To upper bound $\mathcal{R}_n(\mathcal{L})$, we make use of Theorem 1 of Kakade et al. (2008) by observing that $\|w\|_1 \leq \sqrt{t}$. This implies

$$\mathcal{R}_n(\mathcal{L}) \leq \sqrt{\frac{2t\log(2d)}{n}} \max_{x\in T_k} \|x\|_\infty \leq X_k\sqrt{\frac{2t\log(2d)}{n}}, \tag{17}$$

where we recall that X_k is an upper bound of the infinity norm of x in T_k.

Putting together, we have

$$\mathcal{R}_m(\mathcal{F}) \leq 4G_kX_k\sqrt{\frac{2t\log(2d)}{n}} + 2G_k\sqrt{\frac{\max_{(x,x')\in M(T_k)} \zeta^2(x,x')}{n}}. \tag{18}$$

Lastly, we need to upper bound $\left\|f_\zeta^{G_k}\right\|_\infty$ as follows:

$$\left\|f_\zeta^{G_k}\right\|_\infty \leq G_k \cdot |a| + 1 \leq G_k(2\sqrt{t}X_k + Z_k) + 1.$$

By standard uniform convergence via Rademacher complexity, e.g. Lemma 27, we obtain the claimed result. □

C. PAC Guarantee

We aim to show that for any phase k, the distance between the new iterate w_k and the target halfspace w^* is half of that of w_{k-1} and w^*. Thus, after $O(\log \frac{1}{\epsilon})$ iterations, we have an iterate with classification error rate lower than ϵ.

Define the expected hinge loss with respect to clean samples in B_k as

$$L_{\tau_k}(w; D_{X|B_k}) = \mathbb{E}_{x\sim D_{X|B_k}}[\ell_{\tau_k}(w; (x, \text{sign}(w^* \cdot x)))]. \tag{19}$$

Our key tool is a crucial observation from margin-based active learning framework (Balcan et al., 2007; Awasthi et al., 2017; Zhang et al., 2020), which states that in each phase, it suffices to find an iterate w_k whose error rate within the band B_k is a small constant.

We first recall that D_k is the joint distribution D on $\mathcal{X} \times \mathcal{Y}$ conditioned on the event $x \in B_k$. In our analysis, we need the following notion:

$$T_k' := T_k \cap B_k. \tag{20}$$

We will imagine that T_k' is labeled by EX_D^Y. This is only for analysis purpose; in our algorithm T_k' was never involved. Now S_k is a randomly sampled subset of T_k' with size m_k. Again, we remark that the number of labels we need in phase k is m_k.

Due to random sampling with replacement, we have

$$\mathbb{E}_{S_k}[\ell_{\tau_k}(w; S_k)] = \ell_{\tau_k}(w; T_k'). \tag{21}$$

Using standard uniform convergence via Rademacher complexity, we can show that $\ell_{\tau_k}(w; S_k)$ is close to $\ell_{\tau_k}(w; T_k')$. For example, below is implied by Proposition 36 of Shen & Zhang (2021).

Lemma 15. *With probability $1 - \delta'$, the following holds:*

$$\sup_{w\in\mathcal{Q}_k} \left|\ell_{\tau_k}(w; S_k) - \ell_{\tau_k}(w; T_k')\right| \leq \left(1 + \frac{\rho_k X_k}{\tau_k} + \frac{b_k}{\tau_k}\right)\sqrt{\frac{\log(1/\delta')}{m_k}} + \frac{\rho_k X_k}{\tau_k}\sqrt{\frac{2\log(2d)}{m_k}},$$

where $X_k := \max_{x\in T_k'} \|x\|_\infty$.

The following lemma shares the same merit as above, but the expectation is taken over D_k.

Lemma 16 (Lemma 13 of Zhang (2018))**.** *Consider any phase k of Algorithm 2. Let $n'_k := \left|T'_k\right|$. With probability $1-\delta_k$,*

$$\sup_{w\in\mathcal{Q}_k}\left|\ell_{\tau_k}(w;T'_k)-\ell_{\tau_k}(w;D_k)\right| \le K\cdot\log\frac{n'_k d}{\epsilon\delta_k}\cdot\sqrt{\frac{t\log(2d/\delta_k)}{n'_k}}.$$

Lastly, it was shown that as long as the adversarial noise rate is $O(\epsilon)$, the expected hinge loss over the noisy joint distribution D_k is a good proxy of the expected hinge loss over the correctly labeled instances in B_k. This was originally discovered by Awasthi et al. (2017).

Lemma 17 (Lemma 3.8 of Awasthi et al. (2017))**.** *Consider any phase k of Algorithm 2. The following holds for arbitrarily small constant $\kappa>0$, provided that the adversarial noise $\eta\le K\epsilon$ for sufficiently small constant $K>0$:*

$$\sup_{w\in\mathcal{Q}_k}\left|\ell_{\tau_k}(w;D_k)-L_{\tau_k}(w;D_{X|B_k})\right|\le\kappa.$$

Combining Lemma 15, Lemma 16, and Lemma 17, we can show the following useful result.

Proposition 18. *Consider any phase k of Algorithm 2. With probability $1-\delta_k$, the following holds for arbitrarily small constant $\kappa>0$:*

$$\sup_{w\in\mathcal{Q}_k}\left|\ell_{\tau_k}(w;S_k)-L_{\tau_k}(w;D_{X|B_k})\right|\le 3\kappa,$$

provided that $m_k\ge K_1 t\log^3\frac{n_k d}{\delta_k}$ and $n'_k\ge K_1 t\log^4\frac{d}{\epsilon\delta_k}$ for sufficiently large constant $K_1>0$, and the adversarial noise rate $\eta\le K_2\epsilon$ for sufficiently small constant $K_2>0$.

Proof. Due to our hyper-parameter setting, we have $\rho_k=\Theta(\sqrt{t}\tau_k)$ and $b_k=\Theta(\tau_k)$. In addition, using the sub-exponential tail bound of isotropic log-concave distributions, we can show that $\max_{x\in T'_k}\|x\|_\infty\le\max_{x\in T_k}\|x\|_\infty\le O(\log\frac{n_k d}{\delta'})$ holds with probability $1-\delta'$ (see Lemma 26). Thus, by Lemma 15, for any constant $\kappa>0$, when $m_k\ge K_1 t\log^3\frac{n_k d}{\delta_k}$ for large enough constant $K_1>0$, we have that with probability $1-\frac{\delta_k}{3}$,

$$\sup_{w\in\mathcal{Q}_k}\left|\ell_{\tau_k}(w;S_k)-\ell_{\tau_k}(w;T'_k)\right|\le\kappa.$$

In addition, when $n'_k\ge K_1 t\log^4\frac{d}{\epsilon\delta_k}$, Lemma 16 implies that

$$\sup_{w\in\mathcal{Q}_k}\left|\ell_{\tau_k}(w;T'_k)-\ell_{\tau_k}(w;D_k)\right|\le\kappa.$$

The above two inequalities combined with Lemma 17 and the triangle inequality gives the desired result. □

Therefore, we obtain the following key result for the error rate of v_k on $D_{X|B_k}$.

Proposition 19. *Consider any phase k of Algorithm 2. With probability $1-\delta_k$, the following holds for arbitrarily small constant $\kappa>0$:*

$$\Pr_{x\sim D_{X|B_k}}(\mathrm{sign}(v_k\cdot x)\ne\mathrm{sign}(w^*\cdot x))\le 8\kappa,$$

provided that $m_k\ge K_1 t\log^3\frac{n_k d}{\delta_k}$ and $n'_k\ge K_1 t\log^4\frac{d}{\epsilon\delta_k}$ for sufficiently large constant $K_1>0$, and the adversarial noise rate $\eta\le K_2\epsilon$ for sufficiently small constant $K_2>0$.

Proof. Denote by v_k^* the global optimum of the hinge loss minimization problem of Algorithm 2.

We have

$$\begin{aligned}\Pr_{x\sim D_{X|B_k}}(\mathrm{sign}(v_k\cdot x)\ne\mathrm{sign}(w^*\cdot x)) &\le L_{\tau_k}(v_k;D_{X|B_k})\\ &\overset{(a)}{\le}\ell_{\tau_k}(v_k;S_k)+3\kappa\end{aligned}$$

$$\overset{(b)}{\leq} \ell_{\tau_k}(v_k^*; S_k) + 4\kappa$$
$$\overset{(c)}{\leq} \ell_{\tau_k}(w^*; S_k) + 4\kappa$$
$$\overset{(d)}{\leq} L_{\tau_k}(w^*; D_{X|B_k}) + 7\kappa$$
$$\leq 8\kappa.$$

In the above expression, the first step follows from the fact that the hinge loss is an upper bound of the $0/1$ loss, Steps (a) and (d) follow from Proposition 18, Step (b) follows from the definition of v_k, Step (c) follows from the fact that v_k^* is the global minimizer of $\ell_{\tau_k}(w; S_k)$, and the last step follows from Lemma 3.7 of Awasthi et al. (2017) which states that $L_{\tau_k}(w^*; D_{X|B_k})$ can be made to be an arbitrarily small constant κ provided that $\tau_k = \kappa b_k$. □

Finally, we establish the classification error guarantee of v_k on D_X. This is fairly standard due to Balcan et al. (2007). The only minor difference in our analysis is that we need to incorporate the fairness constraint.

In view of Lemma 23, it is more convenient to show the angle between v_k and w^*.

Proposition 20. *Consider any phase k of Algorithm 2. Assume that $w^* \in W_k$. Then with probability $1 - \delta_k$, $\theta(v_k, w^*) \leq 2^{-k-8}\pi$ provided that $m_k \geq K_1 t \log^3 \frac{n_k d}{\delta_k}$ and $n'_k \geq K_1 t \log^4 \frac{d}{\epsilon\delta_k}$ for sufficiently large constant $K_1 > 0$, and the adversarial noise rate $\eta \leq K_2\epsilon$ for sufficiently small constant $K_2 > 0$.*

Proof. For $k = 1$, note that $w^* \in W_1$ and we draw samples from D_X. Thus, Proposition 19 implies

$$\Pr_{x\sim D_X}(\operatorname{sign}(v_k \cdot x) \neq \operatorname{sign}(w^* \cdot x)) \leq 8\kappa. \tag{22}$$

Lemma 23 tells

$$\theta(v_k, w^*) \leq 8\kappa c_2 \leq 2^{-9}\pi, \tag{23}$$

due to the setting of κ.

For any $k \geq 2$, we have

$$\begin{aligned}\Pr_{x\sim D_X}(\operatorname{sign}(v_k \cdot x) \neq \operatorname{sign}(w^* \cdot x),\ x \in B_k) &= \Pr_{x\sim D_{X|B_k}}(\operatorname{sign}(v_k \cdot x) \neq \operatorname{sign}(w^* \cdot x)) \cdot \Pr_{x\sim D_X}(x \in B_k) \\ &\leq 8\kappa c_6 b_k = 8\kappa c_6 \bar{c} r_k,\end{aligned} \tag{24}$$

where the inequality follows from Proposition 19 and Lemma 25.

On the other hand,

$$\theta(v_k, w^*) \leq \pi \left\|v_k - w^*\right\|_2 \leq \pi(\left\|v_k - w_{k-1}\right\|_2 + \left\|w^* - w_{k-1}\right\|_2) \leq 2\pi r_k,$$

where the first inequality is a folklore and the last inequality holds as both v_k and w^* are in W_k. Therefore, since we set $b_k = \bar{c} r_k$ with $\bar{c} \geq \frac{8\pi}{c_4}$, we have $b_k \geq \frac{4}{c_4}\theta(v_k, w^*)$. In view of Lemma 24, we have

$$\begin{aligned}\Pr_{x\sim D_X}(\operatorname{sign}(v_k \cdot x) \neq \operatorname{sign}(w^* \cdot x),\ |v_k \cdot x| \geq b_k) &\leq c_3\theta(v_k, w^*)\exp\big(-\frac{c_4 b_k}{2\theta(v_k, w^*)}\big) \leq 2c_3\pi r_k \exp\big(-\frac{c_4 b_k}{4\pi r_k}\big) \\ &= 2c_3\pi r_k \exp\big(-\frac{c_4\bar{c}}{4\pi}\big).\end{aligned}$$

This combined with (24) gives that for any $k \geq 2$,

$$\Pr_{x\sim D_X}(\operatorname{sign}(v_k \cdot x) \neq \operatorname{sign}(w^* \cdot x)) \leq \Big[8\kappa c_6\bar{c} + 2c_3\pi \exp\big(-\frac{c_4\bar{c}}{4\pi}\big)\Big] r_k. \tag{25}$$

Combining (25), Lemma 23, and the setting $r_k = 2^{-k-3}$, we obtain that

$$\theta(v_k, w^*) \leq \Big[\kappa c_6\bar{c} + \frac{c_3\pi}{4}\exp\big(-\frac{c_4\bar{c}}{4\pi}\big)\Big] c_2 \cdot 2^{-k}.$$

Recall that $\bar{c}$ is a constant such that the coefficient of 2^{-k} is less than $\frac{\pi}{2^8}$ (see Appendix A). This completes the proof. □

Proposition 21. *Consider any phase k of Algorithm 2. If $\theta(v_k, w^*) \le 2^{-k-8}\pi$, then $w^* \in W_{k+1}$ with certainty.*

Proof. The proof follows from some standard algebraic calculations. We will use the following fact: let $\mathcal{S}$ be some set, and $\Pi_{\mathcal{S}}(w)$ be the ℓ_2-projection onto $\mathcal{S}$. Suppose $w^* \in \mathcal{S}$. Then for any w, we have

$$\left\|\Pi_{\mathcal{S}}(w) - w^*\right\|_2 \le \left\|\Pi_{\mathcal{S}}(w) - w\right\|_2 + \|w - w^*\|_2 \le 2\|w - w^*\|_2 . \tag{26}$$

We first show that $\|w_k - w^*\|_2 \le r_{k+1}$. Let $\hat{v}_k = v_k / \|v_k\|_2$. we have

$$\begin{aligned}
\|w_k - w^*\|_2 &= \left\| \mathrm{P}_t(v_k) / \left\|\mathrm{P}_t(v_k)\right\|_2 - w^* \right\|_2 \\
&= \left\| \mathrm{P}_t(\hat{v}_k) / \left\|\mathrm{P}_t(\hat{v}_k)\right\|_2 - w^* \right\|_2 \\
&\le 2\left\|\mathrm{P}_t(\hat{v}_k) - w^*\right\|_2 \\
&\le 4\|\hat{v}_k - w^*\|_2 \\
&\le 2^{-k-4} \\
&= r_{k+1},
\end{aligned}$$

where the first and second inequalities use (26), in the third inequality, we use $\|\hat{v}_k - w^*\|_2 = 2\sin\frac{\theta(v_k, w^*)}{2} \le \theta(v_k, w^*) \le 2^{-k-8}\pi \le 2^{-k-6}$. By the sparsity of w_k and w^*, and our choice $\rho_{k+1} = \sqrt{2t}r_{k+1}$, we always have

$$\|w_k - w^*\|_1 \le \sqrt{2t}\|w_k - w^*\|_2 \le \sqrt{2t}r_{k+1} = \rho_{k+1}.$$

Since w^* has unit ℓ_2-norm and is t-sparse, we also have $\|w^*\|_2 \le 1$ and $\|w^*\|_1 \le \sqrt{t}$. Therefore, $w^* \in \mathcal{Q}_{k+1}$. Since we assumed that w^* has zero fairness error, we have $w^* \in \mathcal{M}_{k+1}$. Thus, $w^* \in W_{k+1}$. □

D. PACF Guarantee, Sample Complexity, and Label Complexity

Theorem 22 (Restatement of Theorem 9)**.** *With probability $1-\delta$, Algorithm 2 runs in polynomial time and returns a halfspace $\hat{w}$ such that $\mathrm{err}_D(\hat{w}) \le \epsilon$. In addition, $\hat{w}$ is α-approximately metric-fair. The sample complexity is $O\big(\big(\frac{1}{\alpha^2} + \frac{1}{\epsilon}\big) \cdot t\log^4\frac{d}{\epsilon\delta} \cdot \log\frac{1}{\epsilon}\big)$, and the label complexity is $O(t\log^3\frac{d}{\alpha\delta} \cdot \log d \cdot \log\frac{1}{\epsilon})$.*

Proof. Note that $w^* \in W_1$. For any phase k, it follows from Proposition 20 and Theorem 12 that the following holds simultaneously with probability $1-\delta_k$: $\theta(v_k, w^*) \le 2^{-k-8}\pi$, and v_k is α-approximately metric-fair.

By iteratively applying Proposition 20 and Proposition 21, we have that with probability $1 - \sum_{k=1}^{k_{\max}}\delta_k \ge 1-\delta$, $v_{k_{\max}}$ is such that $\theta(v_{k_{\max}}, w^*) \le 2^{-k_{\max}-8}\pi$, and it is α-approximately metric-fair.

Using Lemma 23,

$$\Pr_{x\sim D_X}\left(\mathrm{sign}(v_{k_{\max}} \cdot x) \ne \mathrm{sign}(w^* \cdot x)\right) \le \frac{1}{c_1}\theta(v_{k_{\max}}, w^*) \le \frac{\pi}{c_1 \cdot 2^{k_{\max}+8}} = \frac{\epsilon}{2},$$

due to the choice of $k_{\max}$. Now using triangle inequality and our assumption that $\mathrm{err}_D(w^*) \le c_0\epsilon$ for sufficiently small constant $c_0 > 0$, we have

$$\Pr_{(x,y)\sim D}(\mathrm{sign}(v_{k_{\max}} \cdot x) \ne y) \le \Pr_{x\sim D_X}(\mathrm{sign}(v_{k_{\max}} \cdot x) \ne \mathrm{sign}(w^* \cdot x)) + \mathrm{err}_D(w^*) \le \frac{\epsilon}{2} + c_0\epsilon \le \epsilon.$$

The classification error guarantee and fairness error guarantee follow in view of $\hat{w} = v_{k_{\max}}$.

Next, we show the sample complexity. For any phase k, to ensure fairness, we required $|T_k| \ge K \cdot \frac{1}{\alpha^2} \cdot t\log^4\frac{d}{\delta_k}$ in Theorem 12. On the other hand, Proposition 20 holds when $\left|T_k'\right| \ge Kt\log^4\frac{d}{\epsilon\delta_k}$ where $T_k' = T_k \cap B_k$. Since the density of the band B_k is $\Theta(b_k)$ (see Lemma 25), by the Chernoff bound, it suffices to set $|T_k| \ge \Omega(\frac{1}{b_k}(\left|T_k'\right| + \log\frac{1}{\delta_k})) \ge \Omega(\frac{1}{b_k} \cdot t\log^4\frac{d}{\epsilon\delta_k})$.

Hence, in order to fulfill both Theorem 12 and Proposition 20, we need $|T_k| \geq \Omega\big((\frac{1}{\alpha^2}+\frac{1}{b_k})t\log^4\frac{d}{\epsilon\delta_k}\big)$. Consequently, the total number of instances needed by Algorithm 2 is

$$\sum_{k=1}^{k_{\max}}|T_k| \geq \Omega\Big(\big(\frac{1}{\alpha^2}+\frac{1}{\epsilon}\big)\cdot t\log^4\frac{d}{\epsilon\delta}\cdot\log\frac{1}{\epsilon}\Big). \tag{27}$$

Lastly, we analyze the label complexity. Note that labels are needed only when we solve the hinge loss minimization problem. Hence, the number of labels in each phase k is m_k, which is required to be $m_k \geq \Omega(t\log^3\frac{n_k d}{\delta_k}) \geq \Omega\big(t\log^3\frac{d}{\alpha\delta_k}\cdot\log d\big)$. Consequently, the total number of labels needed by Algorithm 2 is

$$\sum_{k=1}^{k_{\max}} m_k \geq \Omega\Big(t\log^3\frac{d}{\alpha\delta}\cdot\log d\cdot\log\frac{1}{\epsilon}\Big). \tag{28}$$

The proof is complete. □

E. Auxiliary Lemmas

Throughout this section, we always assume that D_X is isotropic log-concave.

Lemma 23 (Vempala (2010))**.** *There exist constants $c_1, c_2 > 0$ such that the following holds:*

$$c_1\Pr_{x\sim D_X}(\mathrm{sign}(u\cdot x)\neq\mathrm{sign}(v\cdot x)) \leq \theta(u,v) \leq c_2\Pr_{x\sim D_X}(\mathrm{sign}(u\cdot x)\neq\mathrm{sign}(v\cdot x)).$$

Lemma 24 (Theorem 21 of Balcan & Long (2013))**.** *There are absolute constants $c_3, c_4 > 0$ such that the following holds for all isotropic log-concave distributions D_X. Let u and v be two unit vectors in $\mathbb{R}^d$ and assume that $\theta(u,v)=\theta<\pi/2$. Then for any $b \geq \frac{4}{c_4}\theta$, we have*

$$\Pr_{x\sim D_X}(\mathrm{sign}(u\cdot x)\neq\mathrm{sign}(v\cdot x) \text{ and } |v\cdot x|\geq b) \leq c_3\theta\cdot\exp\big(-\frac{c_4 b}{2\theta}\big).$$

Lemma 25 (Lovász & Vempala (2007a))**.** *There exist constants $c_5, c_6 > 0$ such that the following holds. For any unit vector v and positive real number b,*

$$c_5 b \leq \Pr_{x\sim D_X}(|v\cdot x|\leq b) \leq c_6 b.$$

Lemma 26 (Lemma 20 of Awasthi et al. (2016))**.** *Let T be the set of instances drawn from D_X. With probability $1-\delta$, $\max_{x\in T}\|x\|_\infty \leq O(\log\frac{|S|\,d}{\delta})$*

Lemma 27 (Bartlett & Mendelson (2001))**.** *Consider a loss function $\mathcal{L}: \mathcal{Y}\times\mathcal{A}\to[0,1]$ and a dominating cost function $\phi: \mathcal{Y}\times\mathcal{A}\to[0,1]$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{A}$ and let $(X_i,Y_i)_{i=1}^n$ be independently selected according to the probability measure P. Then, for any integer n and any $0<\delta<1$, with probability at least $1-\delta$ over a sample set S of length n, every f in $\mathcal{F}$ satisfies*

$$\mathbb{E}[\mathcal{L}(Y,f(X))] \leq \frac{1}{n}\sum_{(X,Y)\in S}\phi(Y,f(X)) + \mathcal{R}_n(\tilde{\phi}\circ\mathcal{F}) + \sqrt{\frac{8\log(2/\delta)}{n}},$$

where $\tilde{\phi}\circ\mathcal{F} = \{(x,y)\mapsto\phi(y,f(x))-\phi(y,0) : f\in\mathcal{F}\}$ and $\phi(y,a)$ is an upper bound of $\mathcal{L}(y,a)$.