Active Approximately Metric-Fair Learning

Abstract

Existing studies on individual fairness focus on the passive setting and typically require $O(\frac{1}{\varepsilon^2})$ labeled instances to achieve an ε bias budget. In this paper, we build on the elegant Approximately Metric-Fair (AMF) learning framework and propose an active AMF learner that can provably achieve the same budget with only $O(\log \frac{1}{\varepsilon})$ labeled instances. To our knowledge, this is a first and substantial improvement of the existing sample complexity for achieving individual fairness. Through experiments on three data sets, we show the proposed active AMF learner improves fairness on linear and non-linear models more efficiently than its passive counterpart as well as state-of-the-art active learners, while maintaining a comparable accuracy. To facilitate algorithm design and analysis, we also design a provably equivalent form of the approximate metric fairness based on uniform continuity instead of the existing almost Lipschitz continuity.

1 INTRODUCTION

In modern machine learning applications, an important societal concern is fairness of the machine-learned models. One may think that model-made decisions have no discrimination against minority people, but many case studies show otherwise e.g., Feller et al. [2016], Chan and Wang [2018]. As a result, how to attain fairness in machine-learned models becomes an important research problem and the literature has exploded in recent years. See e.g., some latest surveys Chouldechova and Roth [2018], Mehrabi et al. [2021], Pessach and Shmueli [2022] and references therein.

Model fairness has been studied at both group level and individual level Dwork et al. [2012]. Roughly speaking, group fairness requires certain model outputs to have small disparity across different groups of people, while individual fairness requires model output to be similar on similar individuals. This paper focuses on individual fairness.

Individual fairness was initially formalized as the Lipschitz condition of a prediction model Dwork et al. [2012], and later relaxed to a probabilistic and almost Lipschitz condition called approximate metric-fairness with generalization guarantee Yona and Rothblum [2018]. There are many studies addressing different aspects of individual fairness such as how to design a metric Ilvento [2020], Mukherjee et al. [2020], how to achieve individual fairness with limited resource Kim et al. [2018], Bechavod et al. [2020] or how to combine it with group fairness Zemel et al. [2013], Sharifi-Malvajerdi et al. [2019]. The sample efficiency for achieving individual fairness in passive learning is studied in Balashankar and Lees [2019], Shabat et al. [2020].

This paper studies a new aspect of individual fairness. We ask *if one can attain a more efficient sample complexity for achieving individual fairness through active learning?* To our knowledge, all prior studies focus on the passive setting and maintain an $O(\frac{1}{\varepsilon^2})$ sample complexity for bounding (properly defined) individual bias by ε. See Yona and Rothblum [2018], Balashankar and Lees [2019], Shabat et al. [2020] for example. In this paper, we show it can be improved to $O(\log \frac{1}{\varepsilon})$ through active learning.

To begin with, we first present a new form of approximate metric-fairness that relies on uniform continuity, and show it admits a similar generalization guarantee and is equivalent to the original form based on a connection between uniform continuity and *almost* Lipschitz Vanderbei [1991].

Building on that, we design an active approximate metric-fair learner which labels instances that are fairly close to their neighbors but receive fairly different predictions. We then show this learner only takes $O(\log \frac{1}{\varepsilon})$ labeling to bounded individual unfairness by ε with high probability. Our analysis assumes boundness of a specially designed counter approximate metric-fairness coefficient, and we present two examples on how to calculate it.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

At the end, we experiment the proposed active approximate metric-fair learner on three data sets. We observe it improves individual fairness of both linear and non-linear models more efficiently than its passive counterpart as well as state-of-the-art active learners while maintaining a comparable accuracy, achieving an efficient fairness-accuracy trade-off.

The rest of this paper is organized as follows. We review related topics in Section 2 and present the proposed metric-fairness in Section 3. In section 4, we present the proposed active metric-fair learner and its sample complexity. We discuss implementation in Section 5, present experimental results in Section 6, and conclusion in Section 7. Proofs of certain theoretical results are elaborated in the supplementary material attached after the reference list.

2 BACKGROUND

2.1 FAIRNESS IN MACHINE LEARNING

Today, machine-learned models are widely used in sensitive domains like healthcare and hiring, and it is imperative for them to give fair assessment on human candidates. Take hiring as an example, when a model is used to score the qualification of job candidates, for fairness it should give similar scores to similar candidates disregarding their race or gender (i.e. no racial or gender discrimination). In reality, however, many model assessments are considered unfair Feller et al. [2016], Chan and Wang [2018]. This has motivated intensive research on how to attain fairness in machine-learned models e.g. Grgic-Hlaca et al. [2016], Alabi et al. [2018], Grgić-Hlača et al. [2018], Rothblum and Yona [2018], Mozannar et al. [2020], to name a few.

Model fairness has been studied at both group and individual levels Dwork et al. [2012]. In this paper, we focus on individual fairness which, roughly speaking, requires model output to be similar on similar individuals.

Individual fairness is first formalized as the Lipschitz condition of the model Dwork et al. [2012], and later relaxed to a probabilistic and almost Lipschitz condition called Approximate Metric-Fairness (AMF) Yona and Rothblum [2018][1] Many later studies are built on AMF Balashankar and Lees [2019], Bechavod et al. [2020], Kim et al. [2018]. The active fair learner proposed in this paper is also built on AMF, but we present a new and provably equivalent form based on uniform continuity instead of almost Lipschitz.

One research direction in individual fairness is to attain a proper metric for evaluating individual similarity Ilvento [2020], Mukherjee et al. [2020]. In this paper, we assume a metric is given and focus on how to achieve individual fairness *efficiently* through active learning.

[1] In many following discussions, we will use AMF to represent 'approximate metric-fairness' or 'approximately metric-fair'.

Finally, to our knowledge, existing studies on individual fairness focus on the passive setting, where training data are randomly labeled. Their typical sample complexity for achieving individual fairness is $O(\frac{1}{\varepsilon^2})$ Yona and Rothblum [2018], Balashankar and Lees [2019], Shabat et al. [2020]. In this paper, we focus on the active setting, where training data are strategically labeled. We show the proposed active AMF learner admits an $O(\log \frac{1}{\varepsilon})$ sample complexity, which substantially improves the state-of-the-art result.

2.2 ACTIVE LEARNING

Active learning has been extensively studied in the literature Settles [2009], Aggarwal et al. [2014], Hanneke et al. [2014]. Given a supervised learner, active learning assumes labels of the training data are expensive to query and aims to minimize the query cost by strategically labeling a few data for efficiently improving model accuracy. For example, the uncertainty-based strategy labels data with uncertain model predictions, and the query-by-committee strategy labels data receiving disagreed predictions from a committee of models. Active labeling strategies have been successfully applied in many domains and shown to improve model accuracy more efficiently than random labeling Thompson et al. [1999], Warmuth et al. [2003], Liu [2004], Hoi et al. [2006], Abe et al. [2006], Zhao and Hoi [2013].

On the theory side, active labeling can allows one to learn a model with ε error by labeling only $O(\log \frac{1}{\varepsilon})$ instances, and this is more efficient than random labeling which requires $O(\frac{1}{\varepsilon})$ instances to achieve the same error guarantee Dasgupta [2005], Hanneke [2007], Balcan et al. [2010].

We notice that most active labeling strategies are designed for classification model but very few are for regression model Burbidge et al. [2007], Sugiyama and Nakajima [2009], Cai et al. [2013], Yu and Kim [2010]. To our knowledge, the state-of-the-art strategy for regression model is greedy sampling Wu et al. [2019], which labels data that are most different from the already labeled training data in both feature space and label space.

Our study is related to active learning but differs from existing studies in that they focus on improving accuracy for traditional learners, while we focus on improving individual fairness for AMF learners. Although the two problem settings have fundamental difference, the present study is inspired by disagreement-based active learning Hanneke [2007].

2.3 FAIRNESS IN ACTIVE LEARNING

The intersection of fairness and active learning is a fairly new research direction, and existing studies can be roughly grouped into active labeling Anahideh et al. [2020], Sharaf and Daumé III [2020] and adaptive sampling Abernethy

et al. [2020], Shekhar et al. [2021]. This paper considers the active labeling setting, but differs from the existing study as they focus on improving group fairness for standard learner while we focus on improving individual fairness for AMF learner. Besides, we are the first work that shows active learning can improve the sample complexity for individual fairness to $O(\log \frac{1}{\varepsilon})$, which is not presented in prior studies.

3 APPROXIMATE METRIC-FAIRNESS

In this section, we present a new form of approximate metric-fairness (AMF) that relies on uniform continuity show its equivalence to the original. Then we present the corresponding AMF learner with generalization guarantee.

In this paper, we focus on the regression problem. Let X be an instance space equipped with metric d and distribution D. Let H be a class of regression models[2] defined on X. The original form of AMF Yona and Rothblum [2018] is defined based on almost Lipschitz continuity as follows.

Definition 3.1. A model $h \in H$ is said to be (ε, β) approximately metric-fair with respect to d and D if

$$\Pr_{x,x' \sim D}\{|h(x) - h(x')| > d(x, x') + \beta\} \leq \varepsilon. \quad (1)$$

To facilitate algorithm design and analysis, we propose the following new form of AMF based on uniform continuity.

Definition 3.2. A model $h \in H$ is said to be $(\alpha, \beta, \varepsilon)$ approximately metric-fair with respect to d and D if

$$\Pr_{x,x' \sim D}\{d(x, x') \leq \alpha,\ |h(x) - h(x')| > \beta\} \leq \varepsilon. \quad (2)$$

Our following theorem suggests the two forms are equivalent, which is inspired by the fact that uniform continuity is *almost* Lipschitz Vanderbei [1991].

Theorem 3.3. *Fix any $\alpha, \beta > 0$. Any model with a convex domain is (ε, β) approximately metric-fair with respect to d and D if it is $(\alpha, \beta, \varepsilon)$ approximately metric-fair with respect to metric $d' = \frac{\alpha}{2\beta} \cdot d$ and D, and only if it is $(\alpha, 3\beta, \varepsilon)$ approximately metric-fair with respect to d' and D.*

Proof. Let h be a model with a convex domain. Define two sets $\Psi_1(\beta) = \{(x, x') \mid |h(x) - h(x')| \leq d(x, x') + \beta\}$ and $\Psi_2(\alpha, \beta) = \{(x, x') \mid d'(x, x') \leq \alpha \Rightarrow |h(x) - h(x')| \leq \beta\}$, where '$\Rightarrow$' means 'imply'. We first prove

$$\Psi_2(\alpha, \beta) \subseteq \Psi_1(\beta) \subseteq \Psi_2(\alpha, 3\beta). \quad (3)$$

The left relation holds because, for any β, if there exists an α such that $d'(x, x') \leq \alpha$ implies $|h(x) - h(x')| \leq \beta$, then [Vanderbei, 1991, Theorem 1] implies that $|h(x) - h(x')| \leq d(x, x') + \beta$, where $d(x, x') = \frac{2\beta}{\alpha} d'(x, x')$.

[2]We will use the terms 'model' and 'function' interchangeably.

The right relation follows the fact that, if $|h(x) - h(x')| \leq d(x, x') + \beta = \frac{2\beta}{\alpha} d'(x, x') + \beta$, then $d'(x, x') \leq \alpha$ implies $|h(x) - h(x')| \leq \frac{2\beta}{\alpha} \cdot \alpha + \beta = 3\beta$.

Then, by contrapositive, (3) implies

$$\tilde{\Psi}_2(\alpha, \beta) \supseteq \tilde{\Psi}_1(\beta) \supseteq \tilde{\Psi}_2(\alpha, 3\beta), \quad (4)$$

where $\tilde{\Psi}$ denotes the complement of Ψ. This further implies $\Pr\{\tilde{\Psi}_2(\alpha, \beta)\} \geq \Pr\{\tilde{\Psi}_1(\beta)\} \geq \Pr\{\tilde{\Psi}_2(\alpha, 3\beta)\}$, and the theorem follows by the two fairness definitions. □

Theorem 3.3 shows one form of AMF converges to the other as β decreases, thereby establishing an equivalence between the two forms. It also suggests we can achieve one form of AMF through the other. In the rest of this paper, we will design and analyze AMF learners based on Definition 3.2. For conciseness, we will omit the subscripts in $\Pr_{x,x' \sim D}$ whenever they are clear from the context.

3.1 AMF LEARNING WITH PERFORMANCE GENERALIZATION GUARANTEE

In this section, we introduce an AMF learner based on Definition 3.2 and show its generalization guarantee. Define

$$\Delta_{\alpha,\beta}(h) = \Pr\{d(x, x') \leq \alpha, |h(x) - h(x')| > \beta\}. \quad (5)$$

Let S be a sample of $X \times X$ with cardinality m. Then, an estimate of the probability $\Delta_{\alpha,\beta}(h)$ on sample S is

$$\Delta_{\alpha,\beta}(h; S) = \frac{1}{m} \sum_{(x,x') \in S} \mathbb{I}\{d(x, x') \leq \alpha, \quad |h(x) - h(x')| > \beta\}. \quad (6)$$

A natural idea for AMF learning is to return a model h with small $\Delta_{\alpha,\beta}(h; S)$ and hope it could generalize to a small $\Delta_{\alpha,\beta}(h)$. In this paper, we focus on a realizable case where H contains perfect AMF models (ones satisfying $\Delta_{\alpha,\beta}(h) = 0$) and formalize the goal as follows.

Definition 3.4. Given a hypothesis class H, a loss function ℓ and a labeled training set $L = \{(x_1, y_n), \ldots, (x_n, y_n)\}$ where x_i is the i_{th} instance and y_i is its label, an AMF learner returns a model $h \in H$ by solving

$$\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i), \quad \text{s.t. } \Delta_{\alpha,\beta}(h; S) = 0, \quad (7)$$

where $S = \{(x_i, x_j)\}_{i,j=1,\ldots,n}$.

We can show the AMF learner has a similar generalization guarantee as in Yona and Rothblum [2018] using the following lemma. Let $\mathcal{R}_m(\cdot)$ denote the Rademacher complexity of some hypothesis class for sample size m.

Lemma 3.5. *Fix any $t, \beta > 0$. Let $F : X \times X \to \mathbb{R}$ be a hypothesis class induced from H such that $\forall f \in F$, $f(x, x') = \tau_\beta^t(|h(x) - h(x')|)$ where $\tau_\beta^t(z)$ is a piecewise model outputting 1 if $z > \beta + \frac{1}{t}$, outputting 0 if $z \leq \beta$ and $t(z - \beta)$ otherwise. Then $\mathcal{R}_m(F) \leq 8t \cdot \mathcal{R}_m(H)$.*

Proof Sketch. Repeatedly apply the Rademacher complexity property of composite function with Lipschitz condition e.g. [Bartlett and Mendelson, 2002, Theorem 12] on τ_β^t and abs. See the supplementary material for details. □

Our following theorem presents a generalization guarantee of the AMF learner under mild conditions. We always assume instances are sampled i.i.d..

Theorem 3.6. *Fix any $\alpha, \beta, t > 0$. Suppose $\mathcal{R}_m(H) \in O(1/\sqrt{m})$. Any model $h \in H$ returned by the AMF learner satisfies $\Delta_{\alpha,\beta+1/t}(h) \leq \varepsilon$ with probability at least $1 - \delta$ if $m \geq \frac{1}{\varepsilon^2}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$, where m is the number of $(x, x') \in S$ satisfying $d(x, x') \leq \alpha$ and c is a constant inherited from $O(1/\sqrt{m})$.*

Proof Sketch. A challenge in our analysis is an extra $d(x, x') \leq \alpha$ term that cannot be directly removed using the Rademacher complexity property like in Lemma 3.5. To tackle this, we introduce $V = \{(x, x') \in S; d(x, x') \leq \alpha\}$.

We first transform the analysis of joint event $|h(a) - h(b)| > \beta$ and $d(\alpha, \beta) \leq$ to an analysis of single event $|h(a) - h(b)| > \beta$ by narrowing the domain to V. Then, we derive a generalization bound for event $|h(a) - h(b)| > \beta$, by first relaxing its indicator function to the piecewise function defined in Lemma 3.5, applying the standard generalization argument with $\mathcal{R}_m(F)$ e.g., Mohri et al. [2018], and then connecting $\mathcal{R}_m(F)$ to $\mathcal{R}_m(H)$ using Lemma 3.5. At the end, we connect the obtained result for single event $|h(a) - h(b)| > \beta$ back to a result for the joint event $|h(a) - h(b)| > \beta$ and $d(\alpha, \beta) \leq$ and complete the proof. See the supplementary material for details. □

Theorem A.2 implies one can achieve $(\alpha, \beta, \varepsilon)$ AMF with $O(\frac{1}{\varepsilon^2})$ randomly labeled instances, which is consistent with the sample complexity in Yona and Rothblum [2018]. Constant c depends on the hypothesis class e.g., if H is the set of linear models with proper constraints, we can set c to the maximum norm of the instance Shalev-Shwartz and Ben-David [2014]; if H is the set of kernel machines with proper constraints, we can set c to the product of kernel function bound and gram matrix trace Mohri et al. [2018].

4 ACTIVE AMF LEARNING

In this section, we propose an active AMF learner based on Definition 3.4 and prove it has higher sample efficiency.

Our key idea is to label instances that are close to their neighbors but receive different predictions from some hypothesis. We characterize such instances by the set

$$\begin{aligned}\mathcal{C}_{\alpha,\beta}(H) = \{(x, x') \in X \times X; \exists h \in H, \\ d(x, x') \leq \alpha,\ |h(x) - h(x')| > \beta\}.\end{aligned} \tag{8}$$

We then design the a counter AMF coefficient, which will be used to derive the sample complexity.

Definition 4.1. The counter (α, β) AMF coefficient with respect to a hypothesis class H is

$$\xi_{\alpha,\beta} = \sup_{r>0} \frac{\Pr\{(x, x') \in \mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r))\}}{r}, \tag{9}$$

where $\mathcal{B}_{\alpha,\beta}(r) = \{h \in H; \Delta_{\alpha,\beta}(h) \leq r\}$ is the set of hypotheses that are (α, β, r) AMF.

Roughly speaking, the coefficient measures the largest volume of instance pairs that do not contribute to the fairness achievable in a hypothesis class. We could expect it to be smaller if hypotheses are more fair. We will give two examples later. For conciseness, we will omit the subscripts in $\xi_{\alpha,\beta}$ whenever they are clear from the context.

The proposed active AMF learner is shown in Algorithm 1. In each round, it trains model h on the labeled set using the AMF learner, and then labels instances that are close to the training data but receive different predictions from h. It is clear that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(H)$. The fairness coefficients α, β are assumed preset by the problem, and we can stop labeling when a desired AMF is achieved.

Our following theorem shows that, under proper conditions, Algorithm 1 can with high probability return a model satisfying $(\alpha, \beta, \varepsilon)$ AMF through $O(\log\frac{1}{\varepsilon})$ labeling. Again, assume $\mathcal{R}_m(H) \in O(1/\sqrt{m})$.

Theorem 4.2. *Fix any $\alpha, \beta > 0$. If the counter (α, β) AMF coefficient w.r.t. H is bounded, then with probability at least $1 - \delta$, any $h \in H$ returned by Algorithm 1 satisfies $\Delta_{\alpha,\beta}(h) \leq \varepsilon$ after $O(\log\frac{1}{\varepsilon})$ labeling.*

Proof Sketch. Let $V_q = \{h \in H; \Delta_{\alpha,\beta}(h; S_q) = 0\}$ be the set of 'perfect' AMF models at the end of q rounds of labeling. The goal of our analysis is to show that, if we label $k = \frac{1}{4\xi^2}\left(32c/\beta + \sqrt{\frac{1}{2}\log\frac{1}{\delta'}}\right)$ instances in each round, then by the generalization bound in Theorem A.2, there is

$$\Pr\{\mathcal{C}_{\alpha,\beta}(V_{q+1})\} \leq \frac{1}{2}\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}. \tag{10}$$

with high probability. This implies $Q = \log_2\frac{1}{\varepsilon}$ rounds of labeling, which means $Qk \in O(\log\frac{1}{\varepsilon})$ total labeling, suffices for $\Pr\{\mathcal{C}_{\alpha,\beta}(V_{q+1})\} \leq \varepsilon$. Since $\Delta_{\alpha,\beta}(h) \leq \Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}$ for any $h \in V_q$ by definition, the theorem is proved.

Algorithm 1 Active AMF Learning

Input: an initial labeled training set L, an unlabeled set U, a hypothesis class H, number k.

1: **while** stopping criterion is not met **do**

2: Learn a model $h \in H$ based on sample L using the AMF learner in Definition 3.4.

3: Pick an i.i.d. sample of k instances $u \in U$ satisfying

$$\exists u' \in L,\ d(u,u') \le \alpha,\ |h(u) - h(u')| > \beta. \quad (12)$$

4: Label the selected instances. Then add them to sample L, and remove them from sample U.

5: **end while**

Output: model h.

Let & be logic 'AND' and define event

$$I_\alpha^\beta(x,x';h) := d(x,x') \le \alpha \ \&\ |h(x) - h(x')| > \beta. \quad (11)$$

A key to prove (10) is to split the domain of $\Delta_{\alpha,\beta}(h) = \Pr\{I_\alpha^\beta(x,x';h)\}$ for any $h \in V_{q+1}$ into $(x,x') \in \mathcal{C}_{\alpha,\beta}(V_q)$ and $(x,x') \notin \mathcal{C}_{\alpha,\beta}(V_q)$. Probability on the second subdomain is zero, and probability on the first subdomain can be bounded using Theorem A.2 conditioned on the fact that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_q)$. That bound is smaller than $\frac{1}{2\xi}$ by our choice of k and the definition of ξ, therefore implying $V_{q+1} \subseteq \mathcal{B}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)$ and thus $\Pr\{\mathcal{C}_{\alpha,\beta}(V_{q+1})\} \le \Pr\left\{\mathcal{C}_{\alpha,\beta}\left(\mathcal{B}_{\alpha,\beta}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)\right)\right\} \le \xi \cdot \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi} = \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2}$, where the second inequality is by definition. This proves (10) and thus the theorem. □

4.1 THE COUNTER AMF COEFFICIENT

An important factor in our analysis is the counter AMF coefficient. We give two examples on how to calculate it.

Example 4.3. Fix $\alpha, \beta, B > 0$. Let $h_w(x) = w \cdot x$ be a 1-dimensional linear hypothesis defined on $[-B, B]$, and define $H = \{h_w; w \ge 0\}$. Assume instances are uniformly distributed on $[-B, B]$. If $B > \alpha$, then

$$\Delta_{\alpha,\beta}(h) = \begin{cases} 0, & \text{if } w < \frac{\beta}{\alpha} \\ 1 - \frac{\beta}{\alpha w} + \frac{\beta}{Bw} \ln \frac{\beta}{\alpha w}, & \text{if } w \ge \frac{\beta}{\alpha} \end{cases}, \quad (13)$$

and the counter (α, β) AMF coefficient w.r.t. to H is 1.

Proof. The roadmap of this proof is as follows. We will first derive $\Delta_{\alpha,\beta}(h_w)$ through case study and show it is non-decreasing with respect to w. Based on this, we will then argue the probability in (9) is equivalent to

$$\mathbb{P}_* := \Pr\{|h_*(x) - h_*(x')| > \beta, d(x,x') \le \alpha\}, \quad (14)$$

where h_* is the model satisfying $\Delta_{\alpha,\beta}(h_*) = r$, hence $\xi = \sup_{r>0} \mathbb{P}_*/r = \sup_{r>0} r/r = 1$.

Now we show the detailed proof. For conciseness, we will write h for h_w but with the mind that each h is associated with a w. Also, recall the event notation $I_\alpha^\beta(x,x';h) := d(x,x') \le \alpha \ \&\ |h(x) - h(x')| > \beta$.

Step 1: Characterize $\Delta_{\alpha,\beta}(h)$ for any $h \in H$.

Fix any h. Consider two cases.

(i) If $\alpha w < \beta$, simple geometric analysis shows that event $I_\alpha^\beta(x,x';h)$ is always false so $\Pr_{x,x'}\{I_\alpha^\beta(x,x';h)\} = 0$.

(ii) If $\alpha w \ge \beta$ (which implies $w \ne 0$), then $\alpha \ge \beta/w$. In this case, we can properly partition the domain and have

$$\begin{aligned} &\Pr_{x,x'}\{I_\alpha^\beta(x,x';h)\} \\ &= \mathbb{E}_{x' \in [-B,B]}\left[\Pr_x\{I_\alpha^\beta(x,x';h)\}\right] \\ &= 2\,\mathbb{E}_{x' \in [0,B]}\left[\Pr_x\{I_\alpha^\beta(x,x';h)\}\right] \\ &= 2\int_{x' \in [0,B-\alpha]} \Pr_x\{I_\alpha^\beta(x,x';h)\} \cdot p(x') \\ &\quad + 2\int_{x' \in (B-\alpha, B-\frac{\beta}{w}]} \Pr_x\{I_\alpha^\beta(x,x';h)\} \cdot p(x') \\ &\quad + 2\int_{x' \in (B-\frac{\beta}{w}, B]} \Pr_x\{I_\alpha^\beta(x,x';h)\} \cdot p(x'), \end{aligned} \quad (15)$$

where $\Pr_x\{I_\alpha^\beta(x,x';h)\}$ is the probability defined for x with a fixed x'. In (15), the first equality is by definition, and the second equality is by the observation that $\Pr_x\{I_\alpha^\beta(x,x';h)\}$ is symmetric on $[-B, B]$ (which will become more clear in later analysis). Note that $p(x') = \frac{1}{2B}$.

Now we study each integral separately.

(ii.a) If $x' \in [0, B-\alpha]$, we can show

$$\Pr_x\{I_\alpha^\beta(x,x';h)\} = 1 - \frac{\beta}{w\alpha}. \quad (16)$$

To verify this, let us first fix $x' = 0$ and identify the set of x in $[-\alpha, \alpha]$ that makes event $I_\alpha^\beta(x,0;h)$ true. This case is illustrated in Figure 1. We see all targeted x fall in $[\beta/w, \alpha]$ and (by symmetry) in $[-\alpha, -\beta/w]$. This implies $\Pr_x\{I_\alpha^\beta(x,0;h)\} = \frac{2\cdot(\alpha-\beta/w)}{2\cdot\alpha} = 1 - \frac{\beta}{\alpha w}$. Since h is linear, the above result applies to all $x' \in [0, B-\alpha]$, which implies (16) and thus the first integral equals to $(1 - \frac{\beta}{w\alpha})(1 - \frac{\alpha}{B})$. Note it is non-negative since $\alpha w \ge \beta$ and $B > \alpha$.

(ii.b) If $x' \in (B-\alpha, B-\frac{\beta}{w}]$, we can show

$$\Pr_x\{I_\alpha^\beta(x,x';h)\} = 1 - \frac{\beta}{w(B-x')}. \quad (17)$$

We can verify this in a similar way as in (ii.a), with additional shift of the origin to x' and constraint $x \le B$. Then, geometric analysis suggests all targeted x fall in

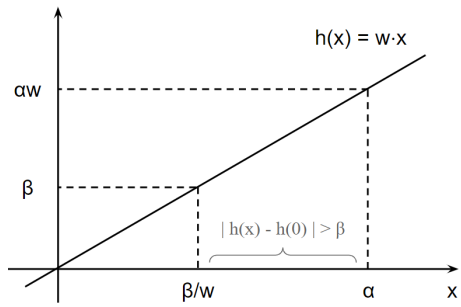


Figure 1

$[x' + \frac{\beta}{w}, B]$ (shorter than interval $[\beta/w, \alpha]$ in Figure 1) and thus $\Pr_x\{I_\alpha^\beta(x, x'; h)\} = \frac{2(B-x'-\beta/w)}{2(B-x')}$. This implies the second integral is $\frac{1}{B}(\alpha - \frac{\beta}{w}(1 - \ln\frac{\beta}{w\alpha}))$. Note it is non-negative as (17) is non-negative by the domain of x'.

(ii.c) If $x' \in (B - \frac{\beta}{w}, B]$, it is easy to see no (x', x) makes event $I_\alpha^\beta(x, x'; h)$ true so $\Pr_x\{I_\alpha^\beta(x, x'; h)\} = 0$. Then the third integral is zero.

Plugging the integrals of (ii.a), (ii.b) and (ii.c) back to (15), and combining results of cases (i) and (ii) gives (13).

Step 2: Show $\Delta_{\alpha,\beta}(h)$ is non-decreasing w.r.t. w.

All we need to show is $\Delta_{\alpha,\beta}(h)$ is non-negative and non-decreasing when $w \geq \frac{\beta}{\alpha}$. The first property is guaranteed since all integrals in (15) are non-negative. To see the second property, take derivative $\frac{\partial \Delta_{\alpha,\beta}(h)}{\partial w} = \frac{\beta\,(B+\alpha(\ln\frac{\alpha w}{\beta}-1))}{w^2\,\alpha\,B}$. Since $w \geq \frac{\beta}{\alpha}$ and $B > \alpha$, we can easily show the derivative is bigger than zero and hence $\Delta_{\alpha,\beta}(h)$ is non-decreasing.

Step 3: Equivalent Probability.

Let $h_* = w_* x$ be the model satisfying $\Delta_{\alpha,\beta}(h_*) = r$. It is not hard to show it exists for every $r \in [0, 1)$ based on (13). Then, results of Step 1 and Step 2 suggest $\mathcal{B}_{\alpha,\beta}(r)$ is the set of linear models satisfying $w \leq w_*$, which implies

$$\Pr\{\mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r))\} = \Pr\{I_\alpha^\beta(x, x'; h_*)\}. \quad (18)$$

To verify this, we first show every $(x, x') \in \mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r))$ makes event $I_\alpha^\beta(x, x'; h_*)$ true. This is true because, for any x, x' with $d(x, x') \leq \alpha$, if there exists an $w \leq w_*$ such that $|wx - wx'| > \beta$, then $|w_* x - w_* x'| \geq |wx - wx'| > \beta$. We then show every (x, x') that makes event $I_\alpha^\beta(x, x'; h_*)$ true is also in $\mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r))$. This is true since h_* exists.

The equivalence implies $\xi = \sup_{r>0} \frac{\Pr\{I_\alpha^\beta(x,x';h_*)\}}{r} = \sup_{r>0} \frac{r}{r} = 1$. The proof is completed. □

The following is another example by similar analysis.

Example 4.4. Fix $\alpha, \beta > 0$. Let $h_z(x) = 1_{x>z}$ be a threshold function defined on $[0, 1]$. Let $H = \{h_z; \alpha \leq z \leq 1 - \alpha\}$. Assume points are uniformly distributed in $[0, 1]$. Then, $\Delta_{\alpha,\beta}(h) = \frac{\alpha(1-\alpha^2)}{1-2\alpha}$ and $\xi \leq \frac{1}{2}$.

5 IMPLEMENTATION

In this section, we discuss three implementation issues.

The first issue is related to the AMF Learner in Definition 3.4. Directly solving (7) is not easy since $\Delta_{\alpha,\beta}(h)$ is non-convex. We propose to approximate the solution by solving

$$\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \lambda\, \tilde{\Delta}_{\alpha,\beta}(h; S), \quad (19)$$

where λ is a regularization coefficient and

$$\tilde{\Delta}_{\alpha,\beta}(h; S) = \frac{1}{n^2\beta^2} \sum_{i,j=1}^{n} M_{ij} \cdot |h(x_i) - h(x_j)|^2, \quad (20)$$

with M being an n-by-n matrix with entries defined as $M_{ij} = \mathbb{I}\{d(x_i, d_j) \leq \alpha\}$. Such approximation can be justified by the following relation.

Lemma 5.1. *Fix any $\alpha, \beta > 0$. We have $\Delta_{\alpha,\beta}(h; S) \leq \tilde{\Delta}_{\alpha,\beta}(h; S)$ for any $h \in S$ and sample S.*

The second issue is related to the base model for AMF learning. We propose to implement a linear model and a kernel regression model approximated by Random Fourier Feature Rahimi and Recht [2007] – we call it 'rff model'.

For the linear model, if instances $x_1, \ldots, x_n \in \mathbb{R}^p$, we can show $\tilde{\Delta}_{\alpha,\beta}(h; S) = \frac{2}{n^2\beta^2} \cdot h^T[x](D - M)[x]^T h$, where $[x]$ is an n-by-p matrix with the i_{th} row being x_i^T. Further, if squared loss is used, then solution to (19) is

$$h = ([x](I - \frac{2\lambda}{n\beta^2}(D - M))[x]^T)^{-1}([x][y]), \quad (21)$$

where $[y] \in \mathbb{R}^n$ is a vector with the i_{th} entry being y_i and D is an n-by-n diagonal matrix with $D_{ii} = \sum_{j=1}^{n} M_{ij}$.

For the rff model, we first calculate random features Rahimi and Recht [2007] and then train a linear model based on them using the AMF learner. Note random features are only used to approximate the prediction model, and we still measure $d(x, x')$ using the original features.

The last issue is related to active learning. Given a labeled training set L and an unlabeled set U, the proposed active AMF learner labels a candidate instance u if there exists $u' \in L$ satisfying $d(u, u') \leq \alpha$ and $|h(u) - h(u')| > \beta$. In principle, we can also pair u with instances in U, as long as the labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_q)$. In practice, pairing u with instances in L is often more efficient (since the label set is often way smaller than the unlabeled set), and leads to slightly better performance as we observe in experiments.

6 EXPERIMENT

We experiment on three data sets. The Insurance data setdat [a] has individual medical costs billed by health insurance

company, and the task is to predict the cost based other attributes. The Life data setdat [b] has the life expectancy in different countries, and the task is to predict the expectancy. We generate the third data set based on public resources. It contains the COVID death rates of 3142 counties in United States and the task is to predict the rate based on other attributes including population density, obesity rate, smoking rate, diabetes rate, elderly population and vaccine rate.

We encode categorical features by dummy variables, address missing data using mean imputation and standardize all features. For higher numerical stability, we re-scale the labels: on the Insurance data set, we divide the medical cost which varies from 4k to 40k by 10k; on the Life data set, we divide the life expectancy which varies from 40 to 90 by 100; on the COVID data set, we multiply the death rate which varies from 0 to 0.01 by 100.

We randomly split each data set into an initial training set (assumed labeled), an unlabeled set (for query) and a testing set. Size of the initial training set is chosen as follows: for the linear base model, it is the feature number on the Insurance and Life data sets, and twice that number on the COVID data set; for the rff base model, it is half of the random feature number. Size of the testing data is 25% of the total data size. The remaining data are treated as unlabeled.

On a data set, we run an active learner for 20 random trials and report the average model performance on the testing sets. Model bias is measured by $\Delta_{\alpha,\beta}(h; S_n)$ defined in (6), with (α, β) set to (2, 0.1) on Insurance, (10, 0.1) on Life and (1.5, 0.001) on COVID. We also experimented with other fairness coefficients and observed similar comparative performance. Model error is measured by the root-mean-squared-error.

We evaluate the proposed active labeling strategy on the linear base model and rff base model respectively, and compare its performance with the following three strategies.

– *Random*: It randomly selects instances to label.

– *Query-by-Committee (QBC)*: It labels instances which receive the largest prediction variance from a committee of models. Following Burbidge et al. [2007], we construct a committee of five models and train each one using a bootstrap sample of the training data, with sample size equals to the training set size divided by the committee size.

– *Uncertainty*: It labels instances which are most different from the training data in both feature space and label space Wu et al. [2019]. To our knowledge, this is a state-of-the-art active labeling method for regression model.

For the metric-fair learner, we pick its regularization coefficient λ that strikes a good balance between fairness and accuracy. Example results of the linear base model on Insurance are shown in Figure 2 (m - n). We see model bias on both training data and testing data decrease as λ increases, which suggests the metric-fair learner can effectively reduce bias and the reduction can generalize (which supports Theorem A.2.) We also see model error first decreases and then increases, exhibiting an overfitting phenomenon. In this case, we will set $\lambda = 1$ as it strikes a good balance. As a result, for the linear base model, λ is set to 1 on Insurance and Life and 0.1 on COVID; for the rff base model, λ is set to 1 on Insurance, 5 on Life and 0.5 on COVID.

For the rff base model, we generate the random features that approximate Gaussian kernel Rahimi and Recht [2007]. The random feature number is set to 100 on Insurance, 400 on Life and 200 on COVID. The gamma coefficient is set to 1e-4 on Insurance, 1e-9 on Life and 1e-2 on COVID. In practice, we observe these configurations lead to good and stable performance of active metric-fair learning.

6.1 RESULTS AND DISCUSSIONS

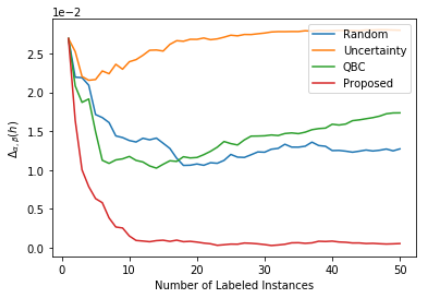
Results of the experimented strategies on both base models across three data sets are shown in Figure 2 (a - l).

In Figure 2 (a-f), we see the proposed active AMF learner reduces model bias more efficiently than other learners, which empirically verifies its efficient sample complexity. We notice it achieves almost zero bias in all cases, supporting our assumption on the realizable case. (And note this is not achieved at the cost of significantly deteriorating accuracy, as explained in the next paragraph.) There seems no consistent pattern on the efficiency of other learners. We notice QBC and uncertainty are often less efficient than random, implying the importance of (efficiently) achieving individual fairness by design, as presented in this study.

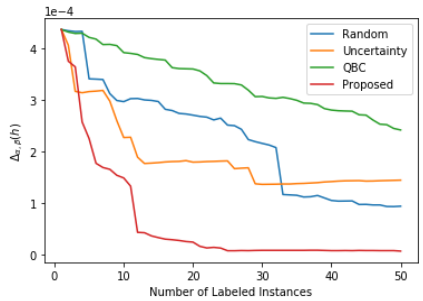
In Figure 2 (g-l), it is not surprising to see that uncertainty based labeling reduces error faster than other strategies. Comparatively, the proposed active AMF learner manages to achieve a comparable reduction rate, suggesting its has an efficient fairness-accuracy trade-off.

7 CONCLUSION

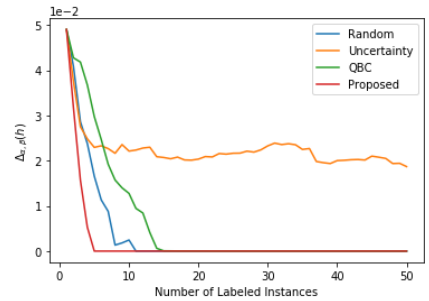
In this paper, we propose the first active approximate metric-fair (AMF) learner and prove it can achieve an ε bias budget by only $\log(\frac{1}{\varepsilon})$ labeled instances with high probability. To our knowledge, this result is a first and substantial improvement over the existing $O(\frac{1}{\varepsilon^2})$ sample complexity for achieving individual fairness by passive learners. Through extensive experiments across three public data sets, we show the proposed active AMF learner improves fairness of two regression models more efficiently than its passive counterpart as well as state-of-the-art active learners, while being able to maintain comparable accuracy, suggesting the proposed learner is efficient in both theory and in practice. Another contribution of this study is to present a provably equivalent form of AMF based on uniform continuity instead of the existing almost Lipschitz continuity.



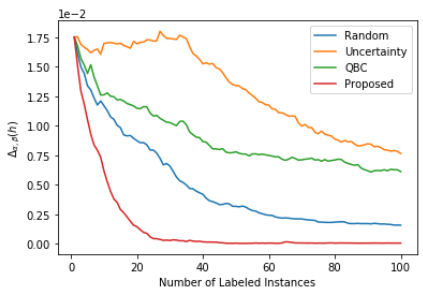
(a) Bias of linear model on Insurance



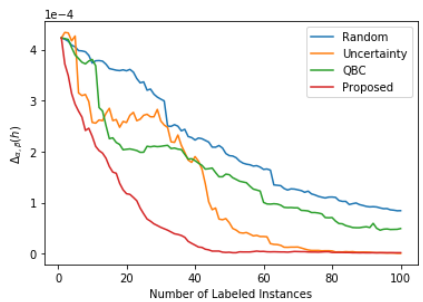
(b) Bias of linear model on Life



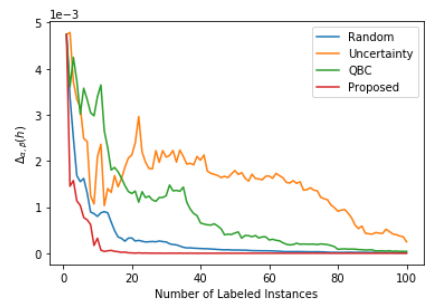
(c) Bias of linear model on COVID



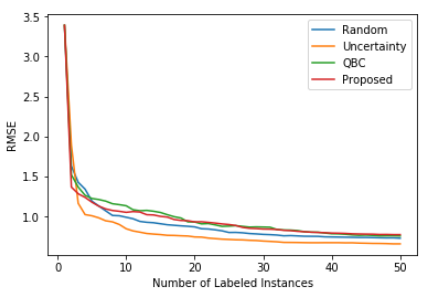
(d) Bias of rff model on Insurance



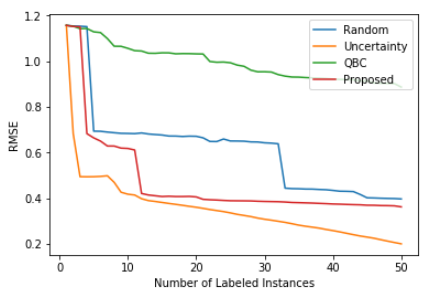
(e) Bias of rff model on Life



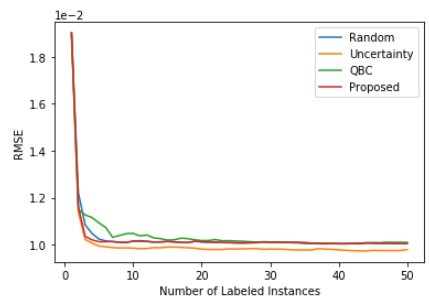
(f) Bias of rff model on COVID



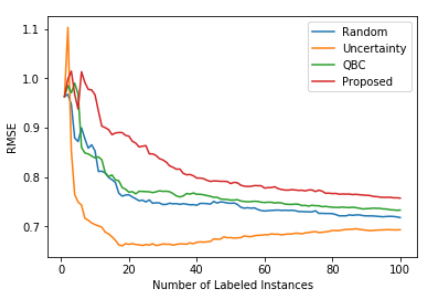
(g) Error of linear model on Insurance



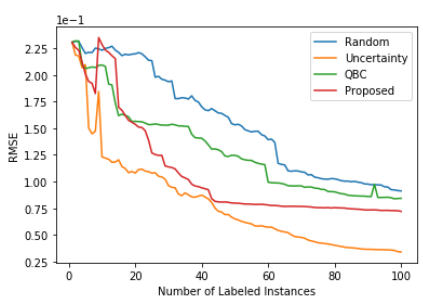
(h) Error of linear model on Life



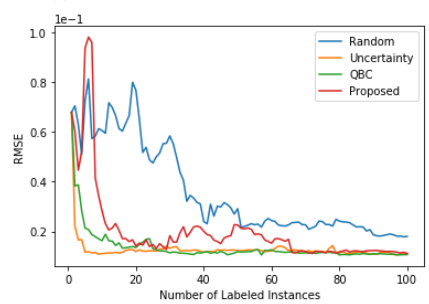
(i) Error of linear model on COVID



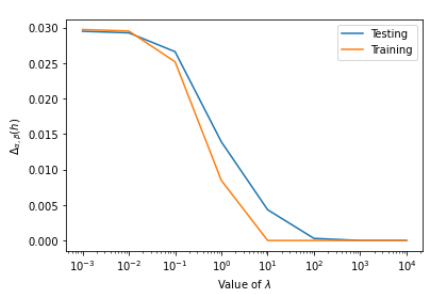
(j) Error of rff model on Insurance



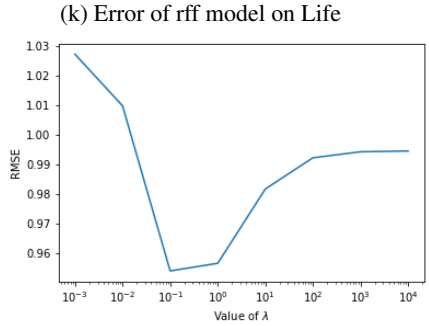
(k) Error of rff model on Life



(l) Error of rff model on COVID



(m) Δ versus λ



(n) RMSE versus λ

Figure 2: Experimental Results

References

a. URL www.kaggle.com/mirichoi0218/insurance.

b. URL www.kaggle.com/kumarajarshi/life-expectancy-who.

Naoki Abe, Bianca Zadrozny, and John Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 504–509, 2006.

Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. *arXiv preprint arXiv:2006.06879*, 2020.

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. Active learning: A survey. In *Data Classification*, pages 599–634. Chapman and Hall/CRC, 2014.

Daniel Alabi, Nicole Immorlica, and Adam Kalai. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*, pages 2043–2066. PMLR, 2018.

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *arXiv preprint arXiv:2001.01796*, 2020.

Ananth Balashankar and Alyssa Lees. Fairness sample complexity and the case for human intervention. *arXiv preprint arXiv:1910.11452*, 2019.

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2):111–139, 2010.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Yahav Bechavod, Christopher Jung, and Steven Z Wu. Metric-free individual fairness in online learning. *Advances in neural information processing systems*, 33: 11214–11225, 2020.

Robert Burbidge, Jem J Rowland, and Ross D King. Active learning for regression based on query by committee. In *International conference on intelligent data engineering and automated learning*, pages 209–218. Springer, 2007.

Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th international conference on data mining*, pages 51–60. IEEE, 2013.

Jason Chan and Jing Wang. Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, 64(7):2973–2994, 2018.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 17, 2016.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 2, 2016.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424, 2006.

Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31, 2018.

Ying Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of chemical information and computer sciences*, 44(6):1936–1941, 2004.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR, 2020.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55 (3):1–44, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Guy N Rothblum and Gal Yona. Probably approximately metric-fair learning. *arXiv preprint arXiv:1803.03242*, 5 (2), 2018.

Burr Settles. Active learning literature survey. 2009.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *Advances in Neural Information Processing Systems*, 33: 13331–13340, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Amr Sharaf and Hal Daumé III. Promoting fairness in learned models by learning to active learn under parity constraints. In *Workshop on Real World Experiment Design and Active Learning. International Conference on Machine Learning*, 2020.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, 32, 2019.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Masashi Sugiyama and Shinichi Nakajima. Pool-based active learning in approximate linear regression. *Machine Learning*, 75(3):249–274, 2009.

Cynthia A Thompson, Mary Elaine Califf, and Raymond J Mooney. Active learning for natural language parsing and information extraction. In *ICML*, pages 406–414. Citeseer, 1999.

R Vanderbei. Uniform continuity is almost lipschitz continuity. Technical report, Technical Report SOR-91-11, Statistics and Operations Research Series . . . , 1991.

Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019.

Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR, 2018.

Hwanjo Yu and Sungchul Kim. Passive sampling for regression. In *2010 IEEE International Conference on Data Mining*, pages 1151–1156. IEEE, 2010.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Peilin Zhao and Steven CH Hoi. Cost-sensitive online active learning with application to malicious url detection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 919–927, 2013.

A SUPPLEMENTARY MATERIAL

Lemma A.1 (Lemma 3.5). *Fix any $t, \beta > 0$. Let $F : X \times X \to \mathbb{R}$ be a hypothesis class induced from H such that $\forall f \in F$, $f(x, x') = \tau_\beta^t(|h(x) - h(x')|)$ where $\tau_\beta^t(z)$ is a piecewise model outputting 1 if $z > \beta + \frac{1}{t}$, outputting 0 if $z \leq \beta$ and $t(z - \beta)$ otherwise. Then $\mathcal{R}_m(F) \leq 8t \cdot \mathcal{R}_m(H)$.*

Proof. Let $G : X \times X \to \mathbb{R}$ be the set of functions induced from h and defined as $\forall g \in G$, $g(a, b) = h(a) - h(b)$. Let abs be the absolute function. Then $f(a, b) = \tau_\beta^t \circ abs \circ g(a, b)$ and we can write, accordingly,

$$F = \tau_\beta^t \circ abs \circ G. \tag{22}$$

We first show $\mathcal{R}_m(F) \leq \mathcal{R}_m(G)$. This is true because

$$\mathcal{R}_m(F) = \mathcal{R}_m(\tau_\beta^t \circ abs \circ G) \leq 2t \cdot \mathcal{R}_m(abs \circ G) \leq 4t \cdot \mathcal{R}_m(G), \tag{23}$$

where both inequalities are by the property of Rademacher complexity for composite function with one component being Lipschitz continuous e.g., [Bartlett and Mendelson, 2002, Theorem 12] and the facts that τ_β^t and abs are both Lipschitz with constants t and 1 respectively.

We then show $\mathcal{R}_m(G) \leq 2 \cdot \mathcal{R}_m(H)$. This is true because

$$\begin{aligned}
\mathcal{R}_m(G) &= \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i g(a_i, b_i) \\
&= \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i [h(a_i) - h(b_i)] \\
&\leq \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(a_i) + \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(b_i) \\
&= 2 \cdot \mathbb{E}_{\{(a_i,b_i)\}}\mathbb{E}_\sigma \sup_{g \in G} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i) \\
&= 2 \cdot \mathcal{R}_m(H),
\end{aligned} \tag{24}$$

where the third equality is based on the fact that σ_i is uniform in {-1, 1} so the expectation with respect to σ_i is the same as the expectation with respect to $-\sigma_i$.

Combining (23) and (24) proves the lemma. □

Theorem A.2 (Theorem 3.6). *Fix any $\alpha, \beta, t > 0$. Suppose $\mathcal{R}_m(H) \in O(1/\sqrt{m})$. Any model $h \in H$ returned by the AMF learner satisfies $\Delta_{\alpha,\beta+1/t}(h) \leq \varepsilon$ with probability at least $1 - \delta$ if $m \geq \frac{1}{\varepsilon^2}\left(16tc + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$, where m is the number of $(x, x') \in S$ satisfying $d(x, x') \leq \alpha$ and c is a constant inherited from $O(1/\sqrt{m})$.*

Proof. To facilitate discussion, define two functions

$$\tau_\beta(z) = \begin{cases} 1, & \text{if } z > \beta \\ 0, & \text{if } z \leq \beta \end{cases}, \tag{25}$$

and

$$\tau_\beta^t(z) = \begin{cases} 1, & \text{if } z > \beta + \frac{1}{t} \\ t(z - \beta), & \text{if } \beta < z \leq \beta + \frac{1}{t} \\ 0, & \text{if } z \leq \beta \end{cases}. \tag{26}$$

By definition, we have

$$\tau_{\beta+\frac{1}{t}}(z) \leq \tau_\beta^t(z) \leq \tau_\beta(z). \tag{27}$$

Recall $S = \{(x_i, x_j)\}_{i,j=1,\ldots,n}$. Let S_α be a subset of S defined as

$$S_\alpha = \{(a, b) \in S \mid d(a, b) \leq \alpha\}. \tag{28}$$

Suppose the size of S_α is m. Then,

$$\begin{aligned}\Delta_{\alpha,\beta}(h;S) &= \frac{1}{n^2}\sum_{i,,j=1}^{n}\mathbb{I}\{|h(x_i)-h(x_j)|>\beta,\ d(x_i,x_j)\le\alpha\}\\ &= \frac{m}{n^2}\cdot\frac{1}{m}\sum_{(a,b)\in S_\alpha}\mathbb{I}\{|h(a)-h(b)|>\beta\}\\ &= \frac{m}{n^2}\cdot\frac{1}{m}\sum_{(a,b)\in S}\tau_\beta(|h(a)-h(b)|).\end{aligned} \tag{29}$$

Recall $F: X\times X\to\mathbb{R}$ is the set of functions induced from τ_β^t and defined as $\forall f\in F$, $f(a,b)=\tau_\beta^t(|h(a)-h(b)|)$. We have that, with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{m}\sum_{(a,b)\in S}\tau_\beta(|h(a)-h(b)|) &\ge \frac{1}{m}\sum_{(a,b)\in S}\tau_\beta^t(|h(a)-h(b)|)\\ &\ge \mathbb{E}[\tau_\beta^t(|h(a)-h(b)|)\mid d(a,b)\le\alpha]-2\mathcal{R}_m(F)-\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\\ &\ge \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|)\mid d(a,b)\le\alpha]-16t\mathcal{R}_m(H)-\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\\ &\ge \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|)\mid d(a,b)\le\alpha]-\frac{1}{\sqrt{m}}\left(16tc+\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right).\end{aligned} \tag{30}$$

where for some constant c. In (30), the first inequality is by (27); the second one is by standard generalization bound[3] with Rademacher complexity e.g. [Mohri et al., 2018, Theorem 3.3] conditioned on $d(a,b)\le\alpha$; the third one is by (27) and Lemma 3.5; and the last one holds since $\mathcal{R}_m\in O(1/\sqrt{m})$. Note the expectation of $(a,b)\in S_\alpha$ in $\mathcal{R}_m\in O(1/\sqrt{m})$ is also conditioned on $d(a,b)\le\alpha$, and we always assume $\mathcal{R}_m\in O(1/\sqrt{m})$ w.r.t. any data proper distribution.

Combining (29) and (30), we see $\Delta_{\alpha,\beta}(h;S)=0$ implies

$$\mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|)\mid d(a,b)\le\alpha]\le\frac{1}{m}\left(16tc+\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right). \tag{31}$$

Further, we can show

$$\Delta_{\alpha,\beta+\frac{1}{t}}(h)\le\mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|)\mid d(a,b)\le\alpha], \tag{32}$$

because

$$\begin{aligned}\Delta_{\alpha,\beta+\frac{1}{t}}(h) &= \int_{(a,b)\in X\times X}\mathbb{I}\{|h(a)-h(b)|>\beta+1/t\}\cdot\mathbb{I}\{d(a,b)\le\alpha\}\cdot p(a,b)\\ &\le \int_{(a,b)\in X\times X}\mathbb{I}\{|h(a)-h(b)|>\beta+1/t\}\cdot p(a,b)\\ &\le \int_{(a,b)\in X\times X}\mathbb{I}\{|h(a)-h(b)|>\beta+1/t\}\cdot p(a,b\mid d(a,b)\le\alpha)\\ &= \mathbb{E}[\tau_{\beta+\frac{1}{t}}(|h(a)-h(b)|)\mid d(a,b)\le\alpha].\end{aligned} \tag{33}$$

Combining (31) and (32), and upper bounding the RHS of (31) by ε implies that $\Delta_{\alpha,\beta+\frac{1}{t}}(h)\le\varepsilon$ whenever

$$m\ge\frac{1}{\varepsilon^2}\left(16tc+\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right). \tag{34}$$

[3] Here we follow Yona and Rothblum [2018] and treat S_α as an i.i.d. sample. If it is not, we can either add an additional constraint that no two pairs in S_α share the same instance so it can be viewed as an i.i.d. sample, or apply a generalization error bound on non-i.i.d. sample e.g. Mohri and Rostamizadeh [2008]. In either case, the order of our result remains the same.

The theorem is proved. □

Theorem A.3 (Theorem 4.2)**.** *Fix any $\alpha, \beta > 0$. Suppose $\mathcal{R}_m(H) \in O(1/\sqrt{m})$ and the counter (α, β) AMF coefficient w.r.t. H is bounded. Then, with probability at least $1-\delta$, any $h \in H$ returned by Algorithm 1 satisfies $\Delta_{\alpha,\beta}(h) \leq \varepsilon$ after $O(\log \frac{1}{\varepsilon})$ labeling.*

Proof. Suppose we have performed q rounds of labeling. Let L_q be the updated training set and S_q be the associated set of instance pairs in Definition 3.4. Define

$$V_q = \{h \in H; \Delta_{\alpha,\beta}(h; S_q) = 0\}. \tag{35}$$

Consider labeling m instances in round $q+1$. First, note that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_q)$ and thus will add to S_q at least m pairs of (x, x') satisfying $d(x, x') \leq \alpha$. Then, by Theorem A.2 and setting $t = 1/\beta$, if $m \geq \frac{1}{4\xi^2}\left(32c/\beta + \sqrt{\frac{1}{2}\log\frac{1}{\delta'}}\right)$, with probability at least $1-\delta'$, any $h \in V_{q+1}$ satisfies

$$\Delta_{\alpha,\beta}(h) \leq 1/(2\xi). \tag{36}$$

Let & be logic 'AND' and define event

$$I_\alpha^\beta(x, x'; h) := d(x, x') \leq \alpha \;\&\; |h(x) - h(x')| > \beta. \tag{37}$$

Then, with probability at least $1-\delta'$, any $h \in V_{q+1}$ satisfies

$$\begin{aligned}
\Pr\{I_\alpha^\beta(x, x'; h)\} &= \Pr\{I_\alpha^\beta(x, x'; h) \,\&\, (x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} + \Pr\{I_\alpha^\beta(x, x'; h) \,\&\, (x, x') \notin \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&= \Pr\{I_\alpha^\beta(x, x'; h) \,\&\, (x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&= \Pr\{I_\alpha^\beta(x, x'; h) \mid (x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \cdot \Pr\{(x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\} \\
&\leq \frac{\Pr\{(x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi},
\end{aligned} \tag{38}$$

where the second equality is by the fact that $\Pr\{I_\alpha^\beta(x, x'; h) \;\&\; (x, x') \notin \mathcal{C}_{\alpha,\beta}(V_q)\} \leq \Pr\{I_\alpha^\beta(x, x'; h) \;\&\; (x, x') \notin \mathcal{C}_{\alpha,\beta}(V_{q+1})\} = 0$, and the inequality is by (36) conditioned on an additional fact that all labeled instances fall in $\mathcal{C}_{\alpha,\beta}(V_{q+1})$. For conciseness, we will write $\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}$ for $\Pr\{(x, x') \in \mathcal{C}_{\alpha,\beta}(V_q)\}$.

Result in (38) implies $V_{q+1} \subseteq \mathcal{B}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)$ and

$$\Pr\{\mathcal{C}_{\alpha,\beta}(V_{q+1})\} \leq \Pr\left\{\mathcal{C}_{\alpha,\beta}\left(\mathcal{B}_{\alpha,\beta}\left(\frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi}\right)\right)\right\} \leq \xi \cdot \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2\xi} = \frac{\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}}{2}, \tag{39}$$

where the first inequality is by the definition of ξ. This result means $\Pr\{\mathcal{C}_{\alpha,\beta}(V_q)\}$ is halved after each round of labeling. Therefore, after $Q := \log_2 \frac{1}{\varepsilon}$ rounds of labeling,

$$\Delta_{\alpha,\beta}(h) \leq \Pr\{\mathcal{C}_{\alpha,\beta}(V_Q)\} \leq \varepsilon, \tag{40}$$

with probability at least $1 - Q\delta'$; where the left inequality is by definition. By then, the total number of labeled instances is $\log_2 \frac{1}{\varepsilon} \cdot \frac{1}{4\xi^2}\left(32c/\beta + \sqrt{\frac{1}{2}\log\frac{1}{\delta'}}\right)$. Setting $\delta = Q\delta'$ and plugging $\delta' = \delta/Q$ in completes the proof. □

Example A.4 (Example 4.4)**.** Fix $\alpha, \beta > 0$. Let $h_z(x) = 1_{x>z}$ be a threshold function defined on $[0, 1]$. Let $H = \{h_z; \alpha \leq z \leq 1-\alpha\}$. Assume points are uniformly distributed in $[0, 1]$. Then, $\Delta_{\alpha,\beta}(h) = \frac{\alpha(1-\alpha^2)}{1-2\alpha}$ and $\xi \leq \frac{1}{2}$.

Proof. We can apply the same proof strategy and show $\Delta_{\alpha,\beta}(h) = \frac{\alpha(1-\alpha^2)}{1-2\alpha} \geq \alpha(1+\alpha)$. (For two points to be within α distance apart, the probability is $(1-2\alpha)2\alpha + 2\alpha(\alpha + \frac{\alpha}{2}) = 1 - \alpha^2$; for z to land in a certain α-length interval within its whole range of length $1-2\alpha$, the probability is $\alpha/(1-2\alpha)$), $\mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r)) = [z-\frac{\alpha}{2}, z] \times [z, z+\frac{\alpha}{2}] \cup [z, z+\frac{\alpha}{2}] \times [z-\frac{\alpha}{2}, z]$, so $\Pr[\mathcal{C}_{\alpha,\beta}(\mathcal{B}_{\alpha,\beta}(r))] = \frac{\alpha^2}{2} \leq \frac{r}{2}$ (because $\alpha + \alpha^2 \leq r$ implies $\alpha^2 \leq r$). Hence $\xi \leq \frac{1}{2}$. □

Lemma A.5 (Lemma 5.1)**.** *Fix any $\alpha, \beta > 0$. We have $\Delta_{\alpha,\beta}(h; S) \leq \tilde{\Delta}_{\alpha,\beta}(h; S)$ for any $h \in S$ and sample S.*

Proof. Since $\mathbb{I}_{x \geq t} \leq \frac{x}{t}$ for any $x, t \geq 0$, we have

$$\begin{aligned}
\mathbb{I}\{d(x_i, x_j) \leq \alpha,\ |h(x_i) - h(x_j)| \geq \beta\} &= \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot \mathbb{I}\{|h(x_i) - h(x_j)|^2 \geq \beta^2\} \\
&\leq \frac{1}{\beta^2} \cdot \mathbb{I}\{d(x_i, x_j) \leq \alpha\} \cdot |h(x_i) - h(x_j)|^2 \\
&= \frac{1}{\beta^2} \cdot M_{ij} \cdot |h(x_i) - h(x_j)|^2.
\end{aligned} \tag{41}$$

Plugging this back to (6) proves the lemma. □