Auditing For Fairness in Machine Learning

November 2018

1 Metric-Free Individual Fairness

1.1 Preliminaries

Audited Classifier The objective of this paper is to audit a classifier $f : \mathfrak{X}^* \to \{0, 1\}$, where $\mathfrak{X}^*$ is a d-dimensional feature space. The Feature space is the Cartesian product of auditing features $\mathfrak{X}$, non-auditing features $\mathfrak{X}^{'}$ and protected attributes $\mathfrak{A}$: $\mathfrak{X}^* = \mathfrak{X} \times \mathfrak{X}^{'} \times \mathfrak{A}$. The data is labeled by $y \in \{0, 1\}$ and the classifier f is trained to label the data $\mathfrak{X}^* \times \{0, 1\}$.

Not all features in $\mathfrak{X}^*$ are to be used to audit the classifier f. The auditor decides on features in $\mathfrak{X}$ along which individuals are considered similar and on features $\mathfrak{X}^{'}$ which should not be considered when comparing two individuals. For example, if f classify loan according to their probability of repayment, the auditor may consider that credit score should be used to define individual similarity, but that zipcode, because correlated with races, should not be an auditing feature, although it was used to learn f. Defining the auditing features empowers the decision maker to choose along which dimensions she wants individuals to be treated the same.

Label Access. The auditor samples $(x^*, f(x^*)$ from a distribution $D \times [-1, 1]$.

Collection of Indicators. An indicator is a subset G of the auditing feature space $\mathfrak{X}$. It defines a function $g : \mathfrak{X} \to \{0, 1\}$ such that $g(x) = 1$ if and only if $x \in G$. $\mathbb{C}$ denotes the collection $\mathbb{C}$ of indicators G. Think of $\mathbb{C}$ as the collection of all groups of individuals for which this paper's notion of fairness treatment will be applied to.

Adjacent distribution From a distribution D over $\mathfrak{X} \times [-1, 1]$ and a value for a protected attribute a, we construct an adjacent distribution $D_a = D|A \neq 1$ with zero probability for all individuals with protected attribute equal to a.

1.2 Metric-Free Individual Fairness.

Metric-Based Individual Fairness. So far, definition of individual fairness in the literature have relied on a similarity metric to impose that similar individuals are treated similarly. With this paper's notations, individual fairness in [2] is defined as follows:

Definition 1.1. *(Individual fairness (from [2])) Let $\delta : \mathfrak{X}^* \times \mathfrak{X}^* \to \mathbb{R}$ be a similarity metric. A classifier f is δ-individually fair if for all $x_1^*, x_2^* \in \mathfrak{X}^*$,*

$$|f(x_1^*) - f(x_2^*)| \leq \delta(x_1^*, x_2^*).$$

Observe, that given its purpose, the similarity metric should only be defined on the space $\mathfrak{X} \times \mathfrak{X}$.For example, in the loan repayment example, the metric will not measure any distance between similar individuals living in different zipcodes.

Metric-Free Individual Fairness. This paper redefines the concept of individual fairness as *individuals with similar features but their protected attributes should be treated the same.*

Definition 1.2. *(Metric-free individual fairness) Consider a α-large collection $\mathbb{C}_\alpha$ of indicators on $\mathfrak{X}$. For $0 \leq \beta < 1$, a classifier f is $(\mathbb{C}_\alpha, \beta)$-metric free individual fair with respect to $\mathfrak{A}$ if for all protected attributes $a \neq a^{'} \in \mathfrak{A}$ and for all $G \in \mathbb{C}_\alpha$:*

$$E_\pi \left| E_{x \sim G}[f(x, .)] - E_{x \sim G_\pi}[f(x, .)] \right| \leq \beta.$$

Metric-free individual fairness formalizes the idea of fairness defined in [1] and borrowed in [2]: *"In a global justice problem, equality of opportunity is satisfied if individual well-being is independent of exogenous irrelevant characteristics"*. The degree of independence to *irrelevant characteristics* is controlled by the value of β: smaller values of β guarantees a stronger level of fairness.

The most noticeable difference between this paper's metric-free definition of individual fairness and previous definitions is that the fairness guarantee is free of similarity metric. The reason is that metric-free fairness guarantees similar treatment across individuals who are exactly similar but along protected attributes. A metric-free definition of individual fairness comes at the cost of no guarantee of similar treatment within protected groups.

Lastly, metric-free individual fairness is a multiple-individuals level notion of fairness. It protects any group of individuals $G \in \mathbb{C}_\alpha$. The collection of indicators $\mathbb{C}_\alpha$ is as in [5] the computational bound on the granularity of metric-free individual fairness. As argued in [5], this relaxation is necessary to audit for fairness in polynomial time. For example, if $\mathbb{C}_\alpha$ is represented by polynomial-sized circuits, the definition of metric-free fairness guarantees fairness within the bound of any indicator that can be computed with polynomial-sized circuits.

1.3 Relation to Differential Fairness

To understand the fairness guarantee offer by metric-free individual fairness, this section relates metric-free individual fairness to a notion of differential individual fairness. Differential individual fairness is a granular extension of the differential fairness as defined by [3]. Consider a collection of indicators $\mathbb{C}_\alpha$. For any $G \in \mathbb{C}_\alpha$, define an adjacent set G_π that is made of all elements of G for which all protected attributes have been reassigned according to π: $G_\pi = \{(x, x^{'}, \pi(a)) | (x, x^{'}, a) \in G\}$. Differential individual fairness imposes that reassigning protected attributes does not change substantially the level dissimilarity in each group of individuals in $\mathbb{C}$. Formally,

Definition 1.3. *(Differential Individual Fairness) f is $(\mathbb{C}_\alpha, \tau)$- differential individually fair if and only if for any reassignment $\pi : \mathfrak{A} \to \mathfrak{A}$ and for any $G \in \mathbb{C}$,*

$$E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_{\pi 1}^*) - f(x_{\pi 2}^*)| \right] \leq E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_1^*) - f(x_2^*)| \right] + \tau$$

Theorem 1.1. *Consider a α-large collection $\mathbb{C}_\alpha$ of indicators on $\mathfrak{X}^*$. Suppose that f is $(\mathbb{C}_\alpha, \beta)$-metric free individual fair with respect to $\mathfrak{A}$. Then, f is $(C_\alpha, 2\beta)$-differential individually fair.*

A classifier could violate similarity constraints as in 1.1 within each protected groups. Theorem 1.1 implies that if f is $(\mathbb{C}_\alpha, \beta)$-metric free individually fair, the degree of violations of the Lipchitz conditions in 1.1 is not increased by more than 2β when reassigning protected attributes with subgroups in $\mathbb{C}_\alpha$. The idea of differentiality is borrowed from the differential privacy work of : metric-free individual fairness guarantees that the degree of dissimilarity within each group of $\mathbb{C}_\alpha$ does not reveal anything about the distribution of protected attributes within that group.

1.4 Relation to Statistical Parity and Equalized Odds

Relation with Statistical Measures of Fairness. The concept of metric-free individual fairness bridges both concepts of statistical fairness and individual fairness. Informally, smaller values of α provides a more granular definition of fairness; larger values of alpha corresponds more to a group/statistical level definition.

Formally, the next results shows that the definition 1.2 encompasses two prevalent notions of statistical fairness:statistical parity, SP, and equalized odds, EO (see [4]):

Theorem 1.2. *(From metric-free fairness to SP and EO) Consider a classifier $f : \mathbb{X} \to \{0, 1\}$. If f is (α, β)-metric individually fair with $\alpha \leq \min_{a \in \mathbb{A}}\{Pr[f = 1 \& A = a]$, then*

(a) f satisfies $\alpha(1 - \beta)$-statistical parity, i.e for all $a, a^{'} \neq a \in \mathbb{A}$

$$|Pr[f = 1, A = a] - Pr[f = 1, A = a^{'}]| \leq \alpha(1 - \beta)$$

(b) f satisfies $\alpha(1 - \beta)$-equalized odds, i.e for all $a, a^{'} \neq a \in \mathbb{A}$ and $y \in \{0, 1\}$

$$|Pr[f = 1, A = a, Y = y] - Pr[f = 1, A = a^{'}, Y = y]| \leq \alpha(1 - \beta)$$

When $\alpha \to 0$ and/or $\beta \to 1$, the definition of metric-free individual fairness implies notion of exact statistical parity or equalized odds (see [4]).

Relation with Metric-Based Measure of Individual Fairness. The main novelty of the concept of metric-free individual fairness is that it does not require defining a metric in the audit space as in [2] or sampling from a metric as in [5]. Although it is a weaker notion of individual fairness, it guarantees the type of protection intended by stronger definition of individuals fairness. [2] provide three motivations to use of individual fairness over aggregate concepts: subset targeting, self-fulfilling prophecy and reduced utility. Subset targeting occurs, for example, when an advertisement company delivers ads related to mortgage refinancing in the same proportions across demographic groups, but target homeowners... need some context to show that metric-free individual fairness covers situations that motivate individual fairness in the first place. Then formal statement of what is covered by metric-free fairness.

Theorem 1.3. *f is a $(\mathbb{C}, \beta)$-metric free individually fair classifier if and only if f is $(\mathbb{C}, \beta)$- differentially fair.*

Theorem 1.4. *Suppose that f is $(\mathbb{C}, \beta)$-metric free individually fair classifier. Let δ be a metric on $\mathbb{X} \times \mathbb{X}$. Define $\mathfrak{C}(\mathbb{C})$ the collection of comparisons induced by $\mathbb{C}$ as $\mathfrak{C}(\mathbb{C}) = \{c : \mathbb{X} \times \mathbb{X} \to \{0, 1\} | \exists g \in \mathbb{C}$ s.t. $c(x, x^{'}) = g(x)g(x^{'})\}$. Then, f is $(\mathfrak{C}(\mathbb{C}), \tau + \beta, \delta)$ multiple fair if and only if*

does not require a metric in the audit space $\mathbb{Z}$, because similarity between individuals is measured between individuals with the same auditing features z but different protected attributes. This is different from the definition of individual fairness in [2] that measures individuals across all individuals, but requires to define a similarity metric. The following definition of individual fairness is borrowed from [2]:

On one hand, metric-free individual fairness is weaker than individual fairness since it only protects group of individuals of size α and since its protection is only partial (unless $\beta \to 1$). On the other hand, metric-free individual fairness proposes a notion of fairness that is true regardless on how individual similarity is measured.

Example where metric-free individual fairness is the right concept:

2 Auditing with Membership Queries

γ- metric free individual unfairness Violation of metric-free fairness can be written as follows:

Definition 2.1. *A classifier f is γ-metric free individually unfair if and only if there exists an indicator $g \in \mathbb{C}$ and an assignment $\pi : \mathfrak{A} \to \mathfrak{A}$ such that*

$$\langle g, u_\pi \rangle \geq \gamma,$$

where $u_\pi(x) = 2 * |f(x^*) - f(x^*_\pi)| - 1$. Then, g is a γ-unfairness certificate.

The definition 2.1 is equivalent to a violation of metric-free individually fairness in 1.2 with $\gamma = \alpha(1-\beta)$ since

$$\langle g, u_\pi \rangle = E[g(x^*)u_\pi(x^*)] = 2Pr[g(x^*) = 1]E[u_\pi(x)|g(x) = 1] - \rho$$

.

$1 - \beta = \frac{\rho+\gamma}{4\alpha'} + \frac{1}{2}$

Auditing

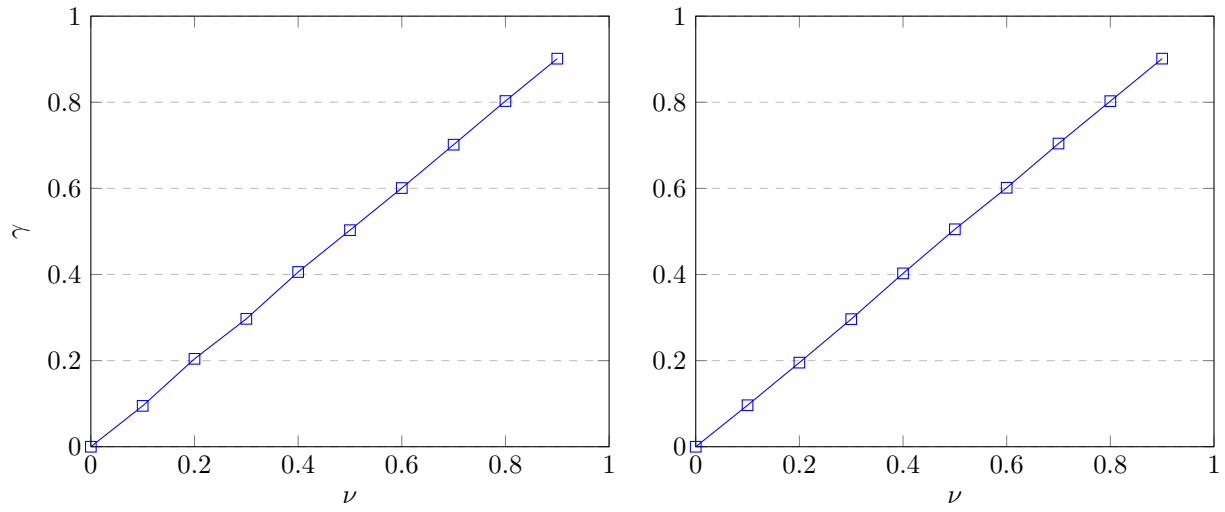
Definition 2.2. *Consider the class of indicators $\mathbb{C}$, a hypothesis class $\mathbb{H}$ and $\gamma' < \gamma$. A $(\mathbb{C}, \mathbb{H}, \gamma, \gamma')$-auditing algorithm with respect to distribution $\mathfrak{D}$ is an algorithm $\mathfrak{M}$ that for any classifier f, any distribution $D \in \mathfrak{D}$, when given oracle access to f,*

1. *With probability $1 - \delta$, provides a $(\mathbb{H}, \gamma')$- unfairness certificate if f is $(\mathbb{C}, \gamma)$-metric free individually unfair.*
2. *With probability 1, returns "fair" if f is $(\mathbb{C}, \gamma)-$ metric free individually fair.*
3. *Runs in poly$(\frac{1}{\delta})$ including queries to $EX(f, D)$ and membership queries $MQ(f)$.*

Agnostic Learning

Definition 2.3. *A concept class $\mathbb{C}$ is agnostically efficiently learnable under distribution $\mathfrak{D}$ if and only if there exists an algorithm $\mathfrak{M}$ that for all $\epsilon, \delta > 0$ in poly$(\frac{1}{\delta}, \epsilon)$ outputs with probability $1 - \delta$ a function $h \in \mathbb{C}$ such that*

$$< f, h > \geq max_{g \in \mathbb{C}} < f, g > + \epsilon.$$



(a) Auditing features are X_1 and X_2.

(b) Auditing features are X_1.

Figure 1: Metric-free individual unfairness versus fraction of individuals unfairly treated.

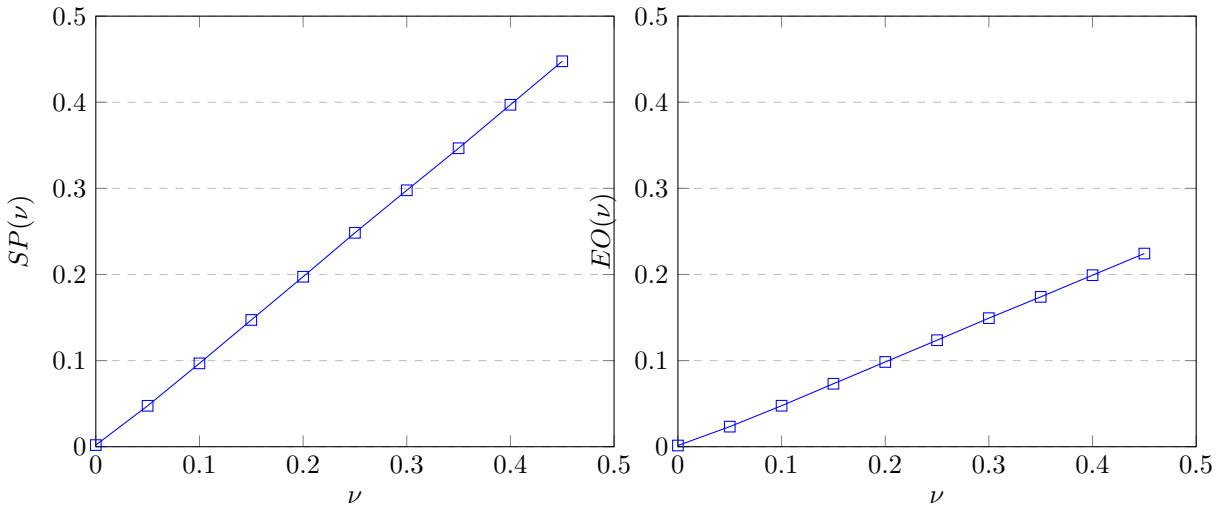
3 Experimental Results

3.1 Synthetic Data

With Oracle Access to Audited Classifier The first set of experiments illustrates how the auditing algorithm in section delivers the correct unfairness certificates when oracle access to the audited classifier f is allowed. Figure 2a shows that when the auditing space $\mathbb{Z}$ is the same as the training features $\mathbb{X}$ used to learn f, the auditing algorithm delivers a γ-unfairness certificate where γ is exactly equal to the fraction of individuals treated unfairly in the sample. Figure 2b shows that the auditing algorithm is robust to using only a subset of the training features, $\mathbb{Z} \neq \mathbb{X}$. Suppose for example that values for X_2 are zipcodes and that the auditor considers that individuals with similar values of X_1 but different zipcodes should be treated similarly.

Metric-free Individual Fairness versus Other Fairness Measures. The first set of experiments illustrates how the concept of metric-free individual fairness relates to other existing definitions of fairness. Figures 2a and 2b plots measures statistical parity and difference in true positive rates across protected groups $A = \{0, 1\}$ for different levels of metric-free individual fairness. Figure 2a shows that the level of statistical parity between protected groups – $SP(\nu) = |Pr[f = 1, A = 0] - Pr[f = 1, A = 1]|$ – is bounded below by ν, which equals $\alpha(1 - \beta)$ in theorem 1.2 and measures the unfairness of classifier f once the data has been modified. Figure 2b illustrates theorem 1.2 for equalized odds: metric-free individual fairness implies that true positive rates – $EO(\nu) = |Pr[f = 1|A = 1, Y = 1] - Pr[f = 1|A = 0, Y = 1]|$ – cannot differ by more than ν across groups $A = 0$ and $A = 1$. Similar results could be obtained for true negative rates.

Figure compares the degree ν of metric-free individual unfairness to the fraction of individual pairs (z, a) and $(z^{'}, a^{'})$ that are treated differently by the classifier f. Theorem indicates that the probability of fair treatment in the sense of [2] should be bounded below by the probability of fair treatment in the sense of this paper.



(a) Statistical parity

(b) True Positive Rates

Overlapping Distributions. The first set of experiments (figure to figure) tests the theoretical results in theorem ... Figure plots the value of the individual fairness measure Δ against the fraction of unfair records ν, when f is a logistic classifier. As stated in theorem, Δ is equal to ν and thus, the plot aligns along the 45° line.

changing the standard deviation σ of the noise ϵ. Figure 2a plots the value of Δ as a function of ν for value of $\sigma \in \{0, 0.1, 0.5, 1\}$ when f is logistic regression and Δ is obtained by training a logistic classifier using auditing features X_1, X_2 and labels $\tilde{R}_f$, where $\tilde{R}_f = R_f$ if $a = 0$ and $\tilde{R}_f = 1 - R_f$ if $a = 0$. The line $\Delta = \nu$ is consistent with theoretical results derived in the previous sections. Moreover, the variance of the noise in Y^* and thus, the accuracy of the classifier f do not affect the experimental results.

4 Appendix

4.1 Proof of Theorem 1.2

Proof. we show the results for statistical parity. The proof is similar for equalized odds. Suppose that f is (α, β)-metric free individually fair with $\alpha \leq \min_{a \in \mathbb{A}}\{Pr[f = 1 \& A = a]$. Let p_a denote the probability than $f(z, a) \neq f(z, a^{'})$ conditional on $f(z, a) = 1$. We first argue that $p_a \leq \alpha(1 - \beta)$. To do so, construct a set $G = \{z \in \mathbb{Z} | f(z, a) = 1 \&\ f(z, a^{'}) = 0\}$. Consider a subset $G^{'}$ of G^c such that $Pr[z \in G^{'}] = \nu - \epsilon$ for some $\epsilon > 0$. We choose ν such that

$$\frac{p_a}{p_a + \nu - \epsilon} = 1 - \beta,$$

or

$$\nu = \epsilon + \frac{\beta}{1 - \beta} p_a. \tag{1}$$

Therefore, $Pr[z \in G^{'} \cup G] = p_a + \nu - \epsilon$. By definition of (α, β)-metric free individual fairness, since $\frac{p_a}{p_a + \nu - \epsilon} = 1 - \beta$, $p_a + \nu - \epsilon < \alpha$. Therefore, by equation (1),

$$p_a < (\alpha - \epsilon)(1 - \beta).$$

Taking the limit $\epsilon \to 0$ leads to $p_a \geq \alpha(1 - \beta)$. The same result holds for $p_{a'} \equiv Pr[f(z, a^{'}) = 1\& \ f(z, a) = 0]$. Moreover,

$$\begin{aligned} Pr[f(z,a) = 1] - Pr[f(z,a^{'}) = 1] &= Pr[f(z,a) = 1 \& f(z,a^{'}) = 0] - Pr[f(z,a) = 0 \& f(z,a^{'}) = 1] \\ &= p_a - p_{a'} \end{aligned} \tag{2}$$

Therefore,

$$|Pr[f(z,a) = 1] - Pr[f(z,a^{'}) = 1]| \leq \alpha(1 - \beta).$$

□

4.2 Proof of theorem 1.1

Suppose that f is a $(\mathbb{C}_\alpha, \beta)$-metric free individually fair classifier. Let $\pi : \mathbb{A} \to \mathbb{A}$ be a reassignment of protected attributes.

$$\begin{aligned} E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_{\pi 1}^*) - f(x_{\pi 2}^*)| \right] &= E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_{\pi 1}^*) - f(x_1^*) + f(x_1^*) - f(x_2^*)) \right. \\ &+ \left. f(x_2^*) - f(x_{\pi 2}^*)| \right] \\ &\leq 2\beta + E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_1^*) - f(x_2^*)| \right] \end{aligned} \tag{3}$$

by using a triangular inequality and then the definition of metric-free individual fairness. A similar argument can be written to show that

$$E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_1^*) - f(x_2^*)| \right] \leq 2\beta + E_{(x_1^*, x_2^*) \sim G \times G} \left[|f(x_{\pi 1}^*) - f(x_{\pi 2}^*)| \right] \tag{4}$$

Therefore f is $(\mathbb{C}_\alpha, 2\beta)$-differentially individually fair.

Conversely, suppose that f is $(\mathbb{C}_\alpha, \tau)$-differentially individually fair. Let G be an indicator set in $\mathbb{C}_\alpha$. Let π be an assignment from $\mathfrak{A}$ to $\mathfrak{A}$. Fix a protected value $a \in \mathfrak{A}$ and construct the reassignment σ such that for all $u \in \mathfrak{A}$, $\sigma(u) = a$. Therefore,

$$\begin{aligned} E_{x^* \sim G} \left[|f(x^*) - f(x_\pi^*)| \right] &= E_{(x^*, x_2^*) \sim G \times G} \left[|f(x^*) - f(x_\pi^*)| \right] \\ &\leq \tau + E_{(x^*, x_2^*) \sim G \times G} \left[|f(x_\sigma^*) - f(x_{\sigma\pi}^*)| \right] \\ &\leq \tau, \end{aligned} \tag{5}$$

since $x_\sigma^* = x_{\sigma\pi}^*$

4.3 Proof of theorem 1.4

Let $S \in \mathfrak{C}$. There exists $G \in \mathbb{C}$ such that $(x, x^{'}) \in S$ if and only if $x, x^{'} \in G$. For $(x, a), (x^{'}, a^{'}) \sim S$,

$$\left| f(x,a) - f(x^{'}, a^{'}) \right| = \frac{1}{|\mathbb{A}|} \left| \sum_{u \in \mathbb{A}} f(x,a) - f(x,u) + \sum_{u \in \mathbb{A}} f(x^{'}, u) - f(x^{'}, a^{'}) + \sum_{u \in \mathbb{A}} f(x,u) - f(x^{'}, u) \right| \tag{6}$$

Moreover, since f is $(\mathbb{C}, \beta)$-metric free individual fair,

$$\begin{aligned} E_{(x,a),(x',a')\sim S}\left[\frac{1}{|\mathbb{A}|}\left|\sum_{u\in\mathbb{A}} f(x,a) - f(x,u)\right|\right] &= E_{x\sim G}\left[\frac{1}{|\mathbb{A}|}\left|\sum_{u\in\mathbb{A}} f(x,a) - f(x,u)\right|\right] \\ &\leq \frac{1}{|\mathbb{A}|}\sum_{u\in\mathbb{A}} E_{x\sim G}\left[|f(x,a) - f(x,u)|\right] \\ &\leq \beta. \end{aligned} \tag{7}$$

$$\begin{aligned} E_{(x,a),(x',a')\sim S}\left[|f(x,a) - f(x^{'},a^{'})|\right] &\leq 2\beta + \frac{1}{|\mathbb{A}|}\sum_{u\in\mathbb{A}} E_{(x,a),(x',a')\sim S}\left[\left|f(x,u) - f(x^{'},u)\right|\right] \\ &\leq 2\beta + \frac{1}{|\mathbb{A}|}\sum_{u\in\mathbb{A}} E_{(x,x')\sim S}\delta(x,x^{'}) \\ &\leq 2\beta + E_{(x,x')\sim S}\delta(x,x^{'}). \end{aligned} \tag{8}$$

4.4 Proof of PAC Reduction

Suppose first that there exists an auditing algorithm $\mathfrak{M}(\epsilon, \delta)$ which audits in time $poly(1/\delta, 1/epsilon)$ for metric-free individual fairness using a concept class $\mathbb{C}$. Denote (x, b) a draw from P. Denote $c^* \in \mathbb{C}$ an indicator $c^* : \mathfrak{X} \to \{-1, 1\}$ Let D denote a distribution over $\mathfrak{X}$ and $P = \{(x, c^*(x))|x \sim D\}$ the corresponding distribution over $\mathfrak{X} \times [-1, 1]$.

Consider a sample S of m examples of P, with m to be determined later on. Construct f such that

$$f(x,a) = \begin{cases} -c(x) & \text{if } c(x) = 1 \text{ and } (x, c(x)) \in S \\ 1 & \text{otherwise .} \end{cases} \tag{9}$$

Consider the following reassignment of protected attributes: fix $a^{'} \in \mathfrak{A}$ with $a^{'} \neq 0$.

$$\pi(a) = \begin{cases} a^{'} & \text{if } a = 0 \\ 0 & \text{if } a \neq 0. \end{cases} \tag{10}$$

Let U_S denote the uniform distribution on S. Oracle access $EX(f, U_S)$ and local membership access $MQ(f, \pi)$ can be simulated from S. Therefore, the auditing algorithm $\mathfrak{M}(\epsilon, \delta)$ can be applied on U_S to audit the classifier f. Denote $u_\pi(x) = |f(x) - f_\pi(x)| - 1$. Since $c^* = u_\pi$ and $\langle c, u_\pi\rangle = 1$, there exists $h \in \mathbb{H}$ such that

$$\langle h, c\rangle(S) \geq 1 - 2\epsilon/3.$$

Denote d the VC dimension of $\mathbb{H} \cup \mathbb{C}$. Then, by uniform convergence, if

$$m = O\left(\frac{9}{4}\frac{d\log(2d/(3\epsilon)) + \log(1/\delta)}{\epsilon^2}\right),$$

with probability $1 - \delta$

$$|\langle c, b\rangle(S) - \langle c, b\rangle(P)| \leq 2\frac{\epsilon}{3}$$

and

$$|\langle h, b\rangle(S) - \langle h, b\rangle(P)| \leq 2\frac{\epsilon}{3}.$$

Therefore,

$$Pr[h = c^*] = \frac{\langle h, u_\pi\rangle(P) + 1}{2} \geq \frac{\langle h, u_\pi\rangle(S) + 1}{2} - 2\frac{\epsilon}{3} \geq \langle c^*, b\rangle(S) - 2\frac{\epsilon}{3} \geq \langle c^*, b\rangle(P) - \epsilon.$$

Morevoer, h is output in $poly(1/\epsilon, 1/\delta) + O\left(\frac{d\log(d/(\epsilon))+\log(1/\delta)}{\epsilon^2}\right)$.

Conversely, assume that $\mathbb{C}$ is PAC learnable by $\mathbb{H}$ in $poly(1/\epsilon, 1/\delta)$. Let f be a classifier defined on $\mathfrak{X}$ and D be a distribution over $\mathfrak{X}$. Let π be a reassignment of protected attributes. Let u_π denote the function $u_a(x) = f(x, a^{'})$ if $a^{'} = a$ and $u_a(x) = -f(x, a^{'})$ otherwise. Denote P the distribution induced by u_a: $P = \{(x, u_a(x))|x \sim D\}$. Suppose that f is $(\mathbb{C}_\alpha, \beta)$-unfair. Therefore, there exists $c \in \mathbb{C}$ such that $\langle c, u_\pi\rangle(P) \geq \beta\alpha$. Since $\mathbb{C}$ is agnostic learnable by $\mathbb{H}$, there exists $h \in \mathbb{H}$ such that

$$Pr[h = u_a] \geq Pr[c = u_a] - \epsilon$$

and such a h is obtained in $poly(1/\epsilon, 1/\delta)$. Moreover,

$$\begin{aligned}\langle h, u_a\rangle = 2Pr[h = u_a] - 1 \geq & 2Pr[c = u_a] - 1 - 2\epsilon \\ & = 2\frac{\langle c, u_a\rangle + 1}{2} - 1 - 2\epsilon \\ & \geq \alpha\beta - 2\epsilon\end{aligned} \tag{11}$$

Therefore, with probability $1 - \delta$, if f is $\alpha\beta$ unfair, the algorithm $\mathfrak{M}(\delta, \epsilon)$ outputs a $(\alpha\beta - 2\epsilon)$ unfairness certificate.

Suppose now that f is $(\mathbb{C}, \beta\alpha)$ metric free individually fair. Then for all a and all $c \in \mathbb{C}$ with $\langle c, u_a\rangle \leq \beta\alpha$, where is u_a is defined as above. Consider c in $c \in \mathbb{C}$. By agnostic efficient learning, algorithm $\mathfrak{M}(\epsilon, \delta)$ delivers with probability $1 - \delta$ a $h \in \mathbb{H}$ such that $Pr[h = u_a] \geq Pr[c = u_a] - \epsilon$. Therefore,

$$\langle h, u_a\rangle = 2Pr[h = u_a] - 1 \leq 2Pr[c = u_\pi] - 1 + 2\epsilon = \langle c, u_\pi\rangle + 2\epsilon \leq 2\beta - \alpha + 2\epsilon.$$

Therefore, algorithm $\mathfrak{M}(\epsilon, \delta)$ will guarantee with probability $1 - \delta$ that f is $(\alpha - \epsilon, \beta + \epsilon)$ metric-free individually fair.

References

[1] Caterina Calsamiglia. Decentralizing equality of opportunity. *International Economic Review*, 50(1):273–290, 2009.

[2] 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. ACM, 2012.

[3] James Foulds and Shimei Pan. An intersectional definition of fairness. 07 2018.

[4] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[5] Michael P Kim, Omer Reingold, and Guy N Rothblum. Fairness through computationally-bounded awareness. *arXiv preprint arXiv:1803.03239*, 2018.