Auditing for Multi-Differential Fairness of Black Box Classifiers

January 2019

1 Introduction

Machine learning algorithms are more and more used to make support decisions that have adverse consequences on an individual's life: for example, classifiers have supported the judicial system to decide whether a criminal offender is likely to recommit a crime; or lender to determine the default risk of a potential borrower. At issue is whether classifiers are fair in the sense of [6] , that is whether classifiers' outcomes are independent of *exogenous irrelevant characteristics* ([6]) or protected attributes, including race and gender. Abundant examples of classifiers' discrimination can be found in many applications (see [5], [2], [3]). A ProPublica story ([3]) reported that a machine learning based risk assessment tool, COMPASS, assigns higher risk to Afro-American defendants than they actually have; and lower risk to Caucasian defendants.

Contestability is challenging for potential victims of machine learning discrimination because (i) many assessment tools are proprietary and are not always required to be transparent about their functioning; and, (ii) there is, at least in the United States, a paper trail of legal cases that have put the burden on the plaintiff to demonstrate disparate treatment, that is to establish that characteristics irrelevant to the task affect the algorithm's outcomes (e.g. see *Ricci et al. vs DeStefano et al.*[1], *Loomis vs. the State of Wisconsin* [4]). In *Loomis vs. the State of Wisconsin* ([4]) , the Wisconsin Supreme Court rules in favor of the use of COMPASS in recidivism risk assessment because, other among things, the plaintiff "failed to meet his burden of showing that the sentencing court actually relied on gender as a factor in sentencing".

This paper defines a framework – differential fairness – and provides a tool for individuals to contest the outcomes of black boxes classifiers whose design, inputs and validation procedures are unknown. Differential fairness is a guarantee that a classifier's outcome are nearly independent of protected attributes conditional on features relevant to the decision making process. The term "differential" is used to emphasize that in our framework a classifier is fair if the classifier's outcomes are statistically nearly identical for two individuals that differ only by their protected attribute. The concept borrows from the differential privacy literature: individual differential fairness measures how much information is leaked by the classifier's outcomes on the distribution of protected attributes. The main advantage of differential fairness is that by conditioning on the individual's features, the framework allows to measure the causal effect of protected attributes on the classifier's outcomes.

However, because it would require to search through a prohibitively large space of individuals, differential fairness cannot be audited for at the individual level. To achieve our goal of auditing

for disparate treatment, we first relax the definition of differential fairness to a notion of multi-differential fairness that guarantees that the classifier's outcomes do not leak any information about the distribution of protected attributes for all sub-populations in a rich collection of groups of individuals.

We show that the relaxation allows to efficiently obtain three fairness diagnostics. First, we reduce the problem of certifying the lack of multi-differential fairness into a learning problem to predict for which individuals a binary protected attribute coincides with the classifier's outcome. Secondly, we propose an algorithm to efficiently find the sub-population for which the classifier violates the most differential fairness. Lastly, we

Contributions Our contributions are as follows:

- We introduce a concept of differential fairness to measure how a classifier leaks information related an individual's protected attribute.
- With a relaxed definition of differential fairness, we reduce the problem of certifying for the lack of differential fairness to the problem of predicting when a binary attribute contribute coincides with the classifier's outcome. This reduction allows to efficiently assess whether the exists a sub-population for which the classifier leaks information related to the distribution of its protected attribute.
-
-

In this paper, we offer and an auditing tool –[I need a name] – that rigorously estimates with some statistical confidence whether a classifier's outcome would differ if the individual had different protected attributes. Differential fairness is a guarantee that a classifier's outcome is indistinguishable whether

would have been different Algorithms-based assessment tools has been widely used because they were thought to be less discriminatory than even the most well meaning human being.

The question is how much additional harm does a black box classifier whose design, data and validation procedures are unknown. If the models have been tinkered with to obtain specific results or maliciously or inadvertently. Notion of constability (Hirsch 2017) and ability to reason about the classifier's outcomes. Different impact versus different treatment? In Ricci vs Stefano, the Suppreme upheld the results of the tests. Are we looking at different treatment instead of different impacts?

Disparate impacts paper: they estimate whether we can predict protected attributes using data D. We argue for a disparate treatment measure that compare the additional disparate impacts caused by the classifier's outcome.

Ricci and Destefano: " liable for disparate impact discrimination only if the exams at issue were not job related and consistent with business necessity, ..." can we link this to choosing relevant auditing features, that are related/consistent with the task at play.

Higher priority of disparate treatment: "the City can avoid disparate-impact liability based on the strong basis in evidence that, had it not certified the results, it would have been subject to disparate-treatment liability"

However, despite a growing literature on classifier's fairness, there are at least two limitations to a systematic approach in defining what classifier means [frame this in terms of bounds: a computation bound and a data/social bound]. First, classifiers are trained and/or audited using samples from distribution that are not independent across external characteristics like race or gender. In fact, we show in this paper that for every classifier with non-trivial accuracy, there exists an imbalanced distribution for which the classifier's outcome will depend on protected attributes for some individuals (see also [17]).This forms the first bound to classifier's fairness, a social bound. Secondly, aggregate notion of fairness that guarantees some definition of fairness to hold on average is not sufficient, since it can hide significant unfairness at the individual level. But only sub-populations that can be efficiently computed can be audited for classifier's fairness. This forms a so-called computational bound to fairness.

This paper introduces a notion of individual differential fairness that makes explicit the information-theory bound. Individual differential fairness is a guarantee that a classifier's outcomes are nearly mean-independent of protected attributes conditional on an individual's features. Borrowing from the differential privacy literature, this definition can be interpreted as a privacy guarantee: a classifier's outcome will leak negligible information about the distribution of protected attributes among individuals sharing the same non-protected features.

Notes: race as a social/legal construct... even with identical feature, two individuals identified in different races will have a different experiment. Directly deal with the social embeddedness of race. Power to communities to understand what the algorithm does. Issues with counterfactual: what is the counterfactual? Yourself with white parents? richer parents.....? It stipulates the source of unfairness but it does not help ... Can AI account for social embeddedness?

Fairness: impartial and just treatment without discrimination or favoritism. Bias and fairness is not interchangeable. But bias is a feature of statistical models; fairness is a feature of human value judgments. Better question: when is fairness instead of what is fairness?

Look into influence functions and Mitigating unwanted biases with adversarial learning

Challenges to formal philosophical models: high unfairness on a small population versus spread of little of unfairness?

Practical applications..... and due diligence

2 Individual and Multi-Differential Fairness

2.1 Preliminary

Notations An individual i is defined by a tuple $((x_i, a_i), y_i)$, where $x_i \in \mathfrak{X}$ denotes individual i's audited features; $a_i \in \mathfrak{A}$ denotes her protected attribute; and $y_i \in \{-1, 1\}$ is the classification provided by a black box classifier f. The auditor draw samples $\{((x_i, a_i), y_i)\}_{i=1}^m$ of size m from a distribution D on $\mathfrak{X} \times \mathfrak{A} \times \{-1, 1\}$.

Features in $\mathfrak{X}$ are not necessarily the ones used to train f. First, the auditor may not have access to all features used to train f. Secondly, the auditor may decide to leave deliberately some features used to train f out of $\mathfrak{X}$ because she believes that those features should not be used to define similarity among individuals. For example, if f classifies 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 train f.

Assumptions In our analysis we make the following assumption:

Assumption 1. *For all* $x \in \mathfrak{X}$, $Pr[A|X = x] > 0$.

Assumption 1 guarantees that the distribution of auditing features conditional on protected attributes have common support: there is no $x \in \mathfrak{X}$ that reveals perfectly the individual's protected attribute.

2.2 Individual Differential Fairness

Individual Differential Fairness We define differential fairness as the guarantee that conditional on features relevant to the tasks, a classifier's outcome is nearly independent of protected attributes:

Definition 2.1. *(Individual Differential Fairness) For* $\delta \in [0, 1)$, *a classifier* f *is* $\delta-$ *differential fair if for all* $x \in \mathfrak{X}$ *and all* $a \in \mathfrak{A}$ *and for all* $y \in \{-1, 1\}$

$$e^{-\delta} \leq \frac{Pr[Y = y|A = a, x]}{Pr[Y = y|A \neq a, x]} \leq e^{\delta} \tag{1}$$

The parameter δ controls how much the distribution of the classifier's outcome Y depends on protected attributes A conditional on auditing features x: larger value of δ implies larger leakage. For example, for a classifier that does not satisfy the fairness condition 2.1 with $\delta = \ln(2)$, there exist individuals x that are twice as likely to be classified $y = 1$ if $A = 1$ than if $A = -1$.

Differential Fairness and Distance of Distributions Since the space of auditing features $\mathfrak{X}$ does not necessarily correspond to the one used to trained the classifier f, for any $x \in \mathfrak{X}$, the auditor access samples drawn from two distributions $Y|A = a, x$ and $Y|a \neq a, x$. Individual fairness constraints these two distributions too be nearly identical when distribution similarity is defined by max-divergence. Formally, the max divergence of two distributions P and Q is defined as

$$D_{\infty}(P||Q) = \max_{y \in Y} \ln \left(\frac{Pr[P = y]}{Pr[Q = y]} \right) \tag{2}$$

The fairness condition in 2.1 can be equivalently rewritten in terms of max divergence as:

$$D_{\infty}((A|x, Y)||(A|x)) \leq \delta \tag{3}$$

If a classifier is δ-individual fair, the posterior and prior distribution of protected attribute conditional on auditing features is bounded by δ.

Relation with Differential Privacy There is an analogy between individual differential fairness for classifiers and differential privacy for database queries. Differential privacy as in [9] guarantees that outcomes from a query are not distinguishable when computed on two adjacent databases that differs only by one record. The fairness condition (1) implies that outcomes from a classifier are not distinguishable for individuals that differ only by their protected attributes. The max-divergence equivalence in 3 shows that differential fairness bounds the information leakage caused by Y conditional on what is already leaked by the auditing features x [placeholder: why does the analogy matter? Possibly, (i) merges the field of fairness with privacy, a field where computer science is "more comfortable" with; (ii). Why is there no balance issue in differential privacy?]

Individual Fairness The definition 2.1 is an individual level definition of fairness, since it conditions the information leakage on auditing features x. Compared to the notion of individual fairness in [8], individual differential fairness does not require to explicit a similarity metric. This is important because defining a similarity metric has been the main limitation of applying the concept of individual fairness. The similarity of treatment in differential fairness is defined in a statistical sense as the max-divergence distance between the distributions $Y|(x, A = a)$ and $Y|(x, A \neq a)$. Differential fairness interprets disparate treatment of a classifier f on an individual with auditing features x as a non-negligible distance between $Y|(x, A = a)$ and $Y|(x, A \neq a)$.

Intention in Disparate Treatment Differential fairness is a useful definition of fairness in machine learning because it provides a test to whether the protected attribute A affects in a causal sense the classifier's outcome. This is important because, at least in the United States, there are legal precedents (see [1], [4], Title VII) that require a plaintiff to demonstrate the disparate treatment was intentional. Causality is defined here as the existence of path (either direct or indirect) between the protected attribute and the classifier's outcome that is not blocked by the auditing features x. Therefore, differential fairness sets a framework to establish causation of protected attributes on a classifier's outcome conditional on the auditing feature space.

2.3 Multi-differential fairness

Although useful, the notion of individual differential fairness suffers from one limitation: it cannot be computationally efficiently audited for. Looking for violations of individual differential fairness will require searching over a set of $2^{|\mathfrak{X}|}$ individuals. Moreover, if $\mathfrak{X}$ is rich enough empirically, a sample from a distribution over $\mathfrak{X} \times \mathfrak{A} \times \{-1, 1\}$ has a negligible probability to have two individuals with the same auditing feature x and different protected attributes a.

Therefore, we relax the definition of individual differential fairness and impose differential individual fairness for group of individuals or sub-populations. Formally, $\mathfrak{C}$ denotes a collection of subsets S of $\mathfrak{X}$. The collection $\mathfrak{C}_\alpha$ is α-strong if for $S \in \mathfrak{C}$ and $y \in \{-1, 1\}$, $Pr[Y = y \;\&\; x \in S] \geq \alpha$.

Definition 2.2. *(Multi-Differential Fairness) Consider a α-strong collection $\mathbb{C}_\alpha$ of sub-populations of $\mathfrak{X}$. For $0 \leq \delta$, a classifier f is $(\mathbb{C}_\alpha, \delta)$-multi differential fair with respect to $\mathfrak{A}$ if for all protected attributes $a, a^{'} \in \mathfrak{A}$, $y \in \{-1, 1\}$ and for all $S \in \mathbb{C}_\alpha$:*

$$Pr[Y = y|A = a, S] \leq e^{\delta} Pr[Y = y|A = a^{'}, S] \tag{4}$$

Multi-differential fairness relaxes the notion of differential fairness by protecting sub-populations instead of individuals. Multi-differential fairness guarantees that the outcome of a classifier f is nearly mean-independent of protected attributes within any sub-population $S \in \mathfrak{C}_\alpha$. The parameter δ controls for the amount of information related to protected attributes that the classifier leaks: smaller value of δ means smaller leakage. The fairness condition 4 applies only to subpopulations with $Pr[Y = y \;\&\; x \in S] \geq \alpha$ for $y \in \{-1, 1\}$. This is to avoid trivial cases where $\{x \in S \;\&\; Y = y\}$ is a singleton for some y, which would imply that $\delta = \infty$.

Disparate Treatment versus Disparate Impact It is interesting to compare the definition of multi-differential fairness in 2.2 with previous definitions of fairness that are based on information leaked by the data about the protected attributes. First, if auditing features is empty, then multi-differential fairness devolves into a notion of disparate impact as in [10] or in [7]. Disparate impact is then a particular case of multi-differential fairness where no disparate treatment statement is made.

On the other hand, if auditing features are all the features but protected attributes used to train f, multi-differential fairness is a framework to test whether there exists a sub-population for which there is a direct causal effect of protected attributes on the classifier's outcomes.

Collection of Indicators. The collection of sub-population $\mathbb{C}$ can be equivalently thought as a family of indicators: for each $S \in \mathbb{C}$, there is an indicator $c_S : \mathfrak{X} \to \{-1, 1\}$ such that $c_S(x) = 1$ if and only if $x \in S$. The relaxation of differential fairness to a collection of groups or sub-population is akin to [16], [14] or [13] where $\mathfrak{C}$ is the computational bound on the granularity of their definition of fairness. The richer $\mathbb{C}$, the stronger the fairness guarantee offers by definition 2.2. However, the complexity of $\mathbb{C}$ is limited by the fact that we identify a sub-population S via random samples drawn from a distribution over $\mathfrak{X} \times \mathfrak{A} \times \{-1, 1\}$. The rest of this paper shows that auditing for multi-differential fairness in polynomial time requires to limit the complexity of $\mathbb{C}$. Potential candidates for $\mathbb{C}$ will be the family short-decision trees or the set of conjunctions of constant number of boolean features. Therefore, auditing for multi-differential fairness will not check whether the fairness condition (4) holds for *all* sub-populations of $\mathfrak{X}$, but only check the fairness condition for all sub-populations that can be *efficiently identifiable*.

3 Auditing, Agnostic Learning and PAC learning

The definition of multi-differential fairness requires to verify that in no sub-population $S \in \mathbb{C}$ with $Pr[S] \geq \alpha$, the classifier leaks information about the distribution of protected attributes. If $\mathbb{C}$ is a rich and large class of subsets of the feature space $\mathfrak{X}$, an auditing algorithm linearly dependent on $|\mathbb{C}|$ can be prohibitively expensive. In this section we show that finding an auditing algorithm reduces to agnostic learning of the class of sub-populations $\mathbb{C}$. That is, there is no $\log(\mathbb{C})$ running time auditing algorithm unless $\mathbb{C}$ is efficiently agnostically learnable.

3.1 Certifying (the Lack) Fairness and Agnostic Learning

Auditing for multi-differential fairness consists firstly, in establishing there exists a fairness violation; secondly, in identifying a sub-population S that violates the most the fairness condition in 2.2.

Multi Differential Fairness and Balanced Distribution The fairness condition 2.2 is unchanged if the feature distribution is reweighted, as long as the reweighting scheme does not depend on the classifier's outcome Y. More formally, for any weights $u : \mathfrak{X} \times \mathfrak{A} \to \mathbb{R}$ such that $u(x, a) > 0$ and $E[u] = 1$,

$$Pr[Y|A = a, S] \leq e^{\delta} Pr[Y|A = a^{'}, S] \iff Pr_u[Y|A = a, S] \leq e^{\delta} Pr_u[Y|A = a^{'}, S], \tag{5}$$

the sub-script u indicating that the probability are taken over the reweighted distribution.

Suppose that for any $a \in \mathfrak{A}$, we have oracle access to the importance sampling weight $w_a(x) = \frac{1-P[A=a|x]}{P[A=a|x]}$. For any distribution D_f over $\mathfrak{X} \times \mathfrak{A} \times \{-1, 1\}$ denote D_f^w the corresponding balanced distribution. Note that once reweighted by w_a, for any sub-population $S \in \mathbb{C}$, auditing features does not reveal anything about the distribution of the protected attribute A: $Pr_w[A = a|X, S] = Pr_w[A \neq a|X, S]$. With a balanced distribution, the multi differential fairness condition can be rewritten as follows: for all protected attributes $a \in \mathfrak{A}$, $y \in \{-1, 1\}$ and for all $S \in \mathbb{C}_\alpha$

$$Pr_w[A = a|S, y] \leq \frac{e^\delta}{e^\delta + 1}, \tag{6}$$

where the sub-script w reminds that the distribution D_f^w is balanced. Since the distribution D_f^w induced by f is balanced, auditing features x do not reveal any information on protected attributes and multi-differential fairness can then be interpreted as an upper bound on ability to predict A given the classifier's outcome for any sub-population $S \in \mathbb{C}$ with $Pr_w[S, y] \geq \alpha$ for $y \in \{-1, 1\}$. A violation of $(\mathbb{C}_\alpha, \delta)$- multi differential fairness is a sub-population $S \in \mathbb{C}_\alpha$ such that

$$Pr_w[A = a|S, y] - \frac{1}{2} \geq \frac{e^\delta}{e^\delta + 1} - \frac{1}{2}. \tag{7}$$

Thereofre, a γ-unfairness certificate is a subset $S \in \mathbb{C}$ such that there exists $y \in \{-1, 1\}$ with

$$Pr_w[S, y] \left\{ Pr_w[A = a|S, y] - \frac{1}{2} \right\} \geq \gamma, \tag{8}$$

with $\gamma = \alpha \left(e^\delta/(1 + e^\delta) - 1/2 \right)$. γ is then a measure of multi-differential unfairness that combines the size of the sub-population where a violation exists and the magnitude of the violation. With balanced distribution, certifying multi-differential fairness is akin to searching for γ-unfairness certificate.

Definition 3.1. *(Certifying Multi-Differential Fairness). Let $\gamma, \epsilon > 0$, $\eta \in (0, 1)$ and $\mathbb{C}_\alpha$ be an α-strong collection of sub-populations in $\mathfrak{X}$. An (ϵ, η)- certifying algorithm $M(\epsilon, \delta)$ is an algorithm that for any sample from a distribution D_f induced by a classifier over $\mathfrak{X} \times \mathfrak{A} \times \{-1, 1\}$, outputs a $\gamma - \epsilon$-unfairness certificate with probability $1 - \eta$ whenever f is γ-multi differential unfair; and, certifies fairness with probability $1 - \eta$ whenever f is is γ-multi differential fair.*

Moreover, $M(\epsilon, \delta)$ is an efficient certifying algorithm if it requires poly$(\log(|\mathfrak{C}_\alpha|, \log(1/\eta), 1/\epsilon))$ samples and runs in poly$(\log(|\mathfrak{C}_\alpha|, \log(1/\eta), 1/\epsilon))$.

Searching for γ-unfairness certificate can be formulated as a problem of detecting correlations that can be written as:

$$Pr_w[AY = c_S] \geq 1 - \rho(y) + 4\gamma \tag{9}$$

with $\rho(y) = Pr_w[A = Y]$. Therefore, certifying the lack of multi-differential fairness can be phrased as a weak agnostic learning problem. A concept class $\mathbb{C}$ is agnostically efficiently learnable if and only if for all $\epsilon, \eta > 0$, there exists an algorithm $\mathfrak{M}$ that given access to a distribution $\{x_i, o_i\} \sim D \times \{-1, 1\}$ outputs with probability $1 - \eta$ in *poly*$(\log(|\mathfrak{C}, \log(\frac{1}{\eta}), \epsilon)$ outputs a function $h \in \mathbb{C}$ such that

$$Pr_D[g = h] + \epsilon \geq max_{c \in \mathbb{C}} Pr_D[g = c].$$

We show that if the collection of subpopulation $\mathbb{C}$ admits an efficient agnostic learner, we could use that learner to construct an algorithm certifying multi-differential fairness.

Theorem 3.1. *Let $\epsilon > 0$, $\beta > 0$ and $\mathbb{C} \subset 2^{\mathfrak{X}}$. There exists an efficient (ϵ, η)-auditing algorithm for $\mathbb{C}$ on balanced distributions if and only if $\mathbb{C}$ admits a (ϵ, η) efficient agnostic learner for any balanced distribution over $\mathfrak{A}$.*

The result in theorem 3.1 makes clear that not all sub-population can be efficiently audited for multi-differential fairness. There are many concept classes $\mathbb{C}$ for which agnostic learning is a NP-hard problem, including for any learning methods that outputs a half-space as an hypothesis (see [11]). However, there are classes for which efficient agnostic learners exist (see [15]).

Based on theorem 3.1 and its proof, we convert the certifying algorithm problem into the following empirical loss minimization: for a sample $\{(x_i, a_i), y_i)\}_{i=1}^m \sim D_f^w$, solve

$$opt = \min_{c \in \mathbb{C}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(a_i y_i = c(x_i)\right) \tag{10}$$

and then compute γ as:

$$\gamma = \frac{opt + \hat{\rho} - 1}{4}, \tag{11}$$

where $\hat{\rho}$ is the sample estimate of $Pr_w[A = Y]$.

3.2 Unbalanced Data

The risk minimization problem allows to certify efficiently the lack of multi-differential fairness. The catch is that we assume that we have oracle access to importance sampling weights w. However, most of the time, importance sampling are unobserved and need to be estimated. Moreover, variance of those estimates are known to be large (see). In this section, we propose a noise infusing technique to certify multi-differential fairness without the need to estimate importance sampling weights.

Importance Sampling and Reweighting At issue with unbalanced distribution is that the multi-differential fairness condition in (6) includes for $a \in \mathfrak{A}$ the term $w_S = Pr[A = a|S]/Pr[A \neq a|S]$:

$$Pr_w[A = a|S, y] \leq \frac{e^{\delta} w_S}{e^{\delta} w_S + 1}, \tag{12}$$

Therefore, algorithm 1, when applied to unbalanced distribution, cannot distinguish the case of high value of δ from the case of low value δ but a high value of w_S. The latter situation is the result of unbalance in the data that could result from social, cultural or historical biases; the former is an issue with the classifier f itself that needs to be audited for.

One approach to obtain w is to directly estimate the density $P[A = a|x]$. This idea of importance sampling is used in propensity-score matching methods (see) in the context of counterfactual analysis. However, exact or estimated importance sampling result in large variance in finite sample. In fact, estimating the distribution $P[A = a|x]$ to obtain the weight $w_a(x)$ may be an overkill. Instead, we observe that if u is a weight function,

$$Pr_w[AY = c] \leq Pr_u[AY = c] + MMD(uP_a, P_{a'}) \tag{13}$$

with equality whenever $u = w_a$. MMD is the maximum mean discrepancy between the distribution of features $P(x, A = a)$ weighted by u and the distribution of $P(x, A \neq a)$:

$$MMD(uP_a, P_{a'}) = \left\| E\left[\frac{1}{n_a} \sum_{i, A=a} u(x)\phi(x) - \frac{1}{n_{a'}} \sum_{i, A=a'} \phi(x) \right] \right\|^2 \tag{14}$$

where $\phi : \mathfrak{X} \to \mathfrak{F}$ is a feature map into a feature space $\mathfrak{F}$ (see [12]). In other words, the maximum mean discrepancy between $uP(., A = a)$ and $P(., A \neq a)$ is measured over all function in the reproducing kernel Hilbert space represented by the kernel $k(x, x^{'}) = \langle \phi(x) | \phi(x^{'}) \rangle$.

Therefore to audit for multi-differential fairness, our approach consists into finding $c \in \mathbb{C}$, ϕ and u that minimizes the empirical counterpart of the upper bound in (14):

$$\begin{aligned} L(w, c, \phi) = & \frac{1}{N} \sum_i u_i \mathbb{1}(c(\phi(x_i)) \neq a_i y_i) + Reg(c) \\ & + \left\| \frac{1}{n_a} \sum_{i, A=a} u_i \phi(x_i) - \frac{1}{N - n_a} \sum_{i, A \neq a} \phi(x_i) \right\|^2 + Reg(u) \end{aligned} \quad . \tag{15}$$

In our implementation MMD_{NET}, the feature representation ϕ is learned via a neural network that is then shared with both tasks of minimizing the re-weighted certifying risk and the distributional shift between $uP(., A = a)$ and $P(., A \neq a)$.

Generic uniform convergence argument allows to derive the sample complexity and correctness of our certifying algorithm 1.

Theorem 3.2. *(Sample Complexity and Correctness of Algorithm 1)Let $\epsilon > 0$ and $\eta \in (0, 1)$. Suppose that $\mathbb{C}$ is a concept class of dimension $d(\mathbb{C}) < \infty$. Algorithm 1 is $(\epsilon, \eta)-$ certifying algorithm for samples of size $m \geq m(\epsilon, \eta, d)$, where*

$$m =$$

Algorithm 1 Certifying Algorithm

1: **Input:** $\{((x_i, a_i), y_i)\}_{i=1}^m$, $\mathbb{C} \subset 2^{|\mathfrak{X}|}$, λ_u, λ_c.

2: $\hat{\rho} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(a_i = y_i)$

3: $u^*, \phi^* = argmin \left\| \frac{1}{n_a} \sum_{i, A=a} u_i \phi(x_i) - \frac{1}{N - n_a} \sum_{i, A \neq a} \phi(x_i) \right\|^2 + \lambda_u \|u\|^2$

4: $c^* = argmin_{c \in \mathbb{C}} \frac{1}{m} \sum_{i=1}^{m} u(x) \mathbb{1}\left(a_i y_i = c(\phi(x_i))\right) + \lambda_c Reg(c)$

5: $\hat{\gamma} a \leftarrow \frac{opt + 1 - \hat{\rho}}{4}$

6: **Return** $\hat{\gamma}-$ unfair

3.3 Unfairness Diagnostics: Worst Violation

Algorithm 1 presented above allows to certify whether any black box classifier is multi-differential fair with only $O(\log(|\mathbb{C}|)$ samples. However, it does not identify the sub-population in $\mathbb{C}_\alpha$ with the strongest violation of multi differential fairness (i.e. with the largest value δ in 2.2). This is because algorithm 1 does not distinguish a large sub-population S with low value of δ from a smaller sub-population with larger value of δ. Finding the strongest violation is useful to (i) diagnostic the source of multi-differential unfairness of a classifier; and, (ii) identify the individuals which could be the most harmed by a classifier's outcome.

Worst Violation Problem The objective is to identify for any $a \in \mathfrak{A}$ the sub-population S in $\mathbb{C}$ that solves:

$$\delta_m \equiv \sup_{S \in \mathbb{C}} \log \left(\frac{Pr[A = a|S, y]}{Pr[A \neq a|A, y]} \Big/ \frac{Pr[A = a|S]}{Pr[A \neq a|S]} \right). \tag{16}$$

To illustrate the challenges of the worst-violation problem, consider the case of a binary protected attributes $A = \{-1, 1\}$ where there exist $S_0, S_\delta \in \mathbb{C}$ with no violation of multi differential fairness in sub-population S_0, but a $\delta-$ multi differential fairness violation in S_δ (e.g. $P[Y|A = 1, S_\delta] < e^\delta P[Y|A = -1, S_\delta]$). Consider $c, c^{'} \in \mathbb{C}$ such that $c(x) = 1$ if and only if $x \in S_\delta$ and $c^{'}(x) = 1$ if and only if $x \in S_\delta \cup S_0$. Both c and $c^{'}$ have the same accuracy on $S_\delta \cup S_0$ once the distribution is rebalanced. Therefore, algorithm 1 will pick indifferently c or $c^{'}$ as unfairness certificate, although $c^{'}$ does not single out S_δ as a worst violation.

Worst Violation Algorithm At issue in the previous example is that for the sub-population S_0 with no violation of multi differential fairness, choosing $c = 1$ or $c = -1$ will lead to same empirical risk used in our certifying procedure 1. Our worst violation algorithm 2 puts a slightly larger weight on samples $((x_i, a_i), y_i)$ whenever $a_i y_i \neq 1$ so that the empirical risk is now smaller if $c = -1$ and there is no violation of multi differential fairness. More generally, our worst violation algorithm is an iterative procedure such that at each iteration t, the weight on samples $((x_i, a_i), y_i)$ whenever $a_i y_i \neq 1$ is increased to $1 + \xi t$, with $\xi > 0$ and the solution c_t^* of the empirical risk minimization (10) or (11) identifies a sub-population $S_t = \{x|c_t^*(x) = 1\}$ for which $\delta \geq \log(1 + \xi t)$. The algorithm 2 terminates whenever either $|S| \leq \alpha$ or $c_t^*(x) = -1$ for all x. At the last iteration before termination, theorem 3.3 shows that algorithm 2 will identify a sub-population with a δ-multi differential fairness violation with δ asymptotically close to δ_m.

Theorem 3.3. *Suppose $\nu > 0, \epsilon > 0, \eta \in (0, 1)$ and $\mathbb{C} \subset 2^{\mathfrak{X}}$ is $\alpha-$strong. Denote δ_m the worst violation of multi differential fairness for $\mathbb{C}$ as defined in* (16)*. With probability $1 - \eta$, with $O()$ samples and after $O()$ iterations, algorithm 2 learns $c \in \mathbb{C}$ such that*

$$\log \left(\frac{Pr_w[A = a|y, c(x) = 1]}{Pr_w[A \neq a|y, c(x) = 1]} \right) \geq \delta_m - \epsilon. \tag{17}$$

3.4 Unfairness Diagnostic: Individual Recourse

Although multi-differential fairness is only a sub-population level definition of fairness, in this section, we explore how this framework can inform on the harm made by a classifier's outcome on a given individual. At the individual level, the harm is measured by the max-divergence between the distributions $Y|(A = a, x)$ and $Y|(A \neq a, x)$:

$$\delta(x) \equiv max_{y \in Y} \ln \left(\frac{Pr_w[Y = y|A = a, x]}{Pr_w[Y = y|A \neq a, x]} \right). \tag{18}$$

In this section, we propose a lower bound of $\delta(x)$ that can be efficiently computed for any individual x.

4 Experiments

4.1 Synthetic Data

A synthetic data is constructed by drawing independently two features X_1 and X_2 from two normal distribution $N(0, 1)$. We consider a binary protected attribute $\mathfrak{A} = \{-1, 1\}$ drawn from Bernouilli

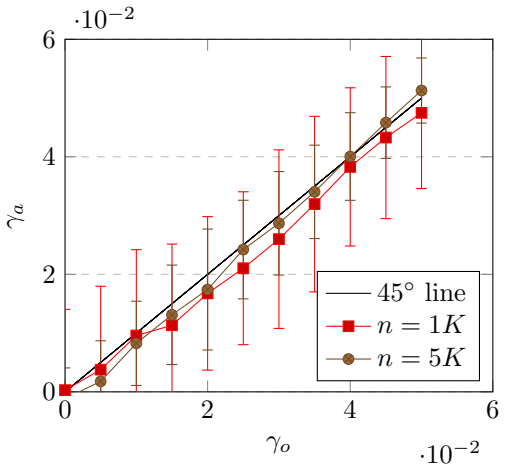
Algorithm 2 Worst Violation Algorithm

Input: $\{((x_i, a_i), y_i)\}_{i=1}^m$, $\mathbb{C} \subset 2^{|\mathfrak{X}|}$, ξ, $alpha$
$\delta_{-1} = 0$; $\delta_0 = 1$, $\alpha_0 = 1$
while $\delta_t >= \delta_{t-1}$ or $\alpha_t > \alpha$ **do**
    for i=1..m **do** $u_{it} \leftarrow u_i(1 + \xi t)$ if $a_i y_i = -1$
    $c^* = argmin_{c \in \mathbb{C}} \frac{1}{m} \sum_{i=1}^{m} u_{it}(x) \mathbb{1}\left(a_i y_i = c(\phi(x_i))\right) + \lambda_c Reg(c)$
    $\hat{\delta}_t \leftarrow \dfrac{\sum_{i=1, c(x_i)=1}^{m} u_i(x_i) \mathbb{1}(y_i = 1) \mathbb{1}(a_i = 1)}{\sum_{i=1, c(x_i=1)}^{m} u_i(x_i) \mathbb{1}(y_i = 1) \mathbb{1}(a_i = -1)}$
    $\hat{\alpha_t} \leftarrow \dfrac{\sum_{i=1)}^{m} u_i(x_i) \mathbb{1}(c_i(x_i) = 1)}{\sum_{i=1}^{m} u_i(x_i)}$
    $t \leftarrow t + 1$
Return $\ln(\delta_t)$.

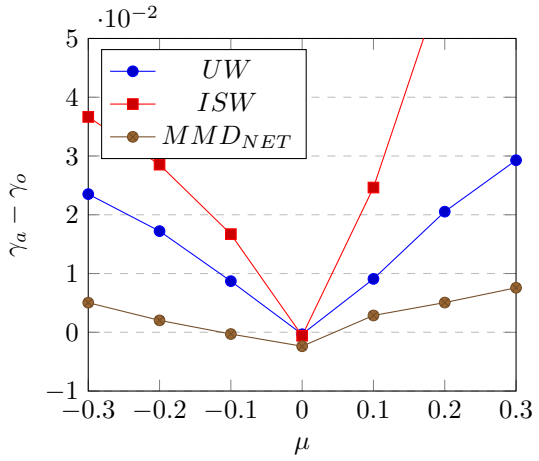
distribution with $A = 1$ obtained with probability $w(x) = \frac{e^{\mu * (x_1 - x_2))^2}}{1 + e^{\mu * (x_1 + x_2))^2}}$ and $A = -1$ with probability $1 - w(x)$ for $x = (x_1, x_2)$. μ is an unbalance factor. $\mu = 0$ means that the data is perfectly balanced with $w(x) = \frac{1}{2}$. As μ increases, the distribution $Pr(x|A = 1)$ becomes more dense in the for points away from the 45° diagonal in a (x_1, x_2) plane. The data is labeled according to the sign of $(X_1 + X_2 + e)^3$, where is e is a noise drawn from $N(0, 0.2)$. The audited classifier f is a logistic regression classifier that is altered to generate instances of differential unfairness: with probability $1 - \nu \in [0, 1)$, the sign of the classifier's outcomes Y from individuals with $A = -1$ is changed from -1 to $+1$ if $x_1^2 + x_2^2 \leq 1$; if $A = 1$, all classifier's outcomes are changed from -1 to $+1$ if $x_1^2 + x_2^2 \leq 1$. For $\nu = 0$, the audited classifier is differentially fair; however, as ν increases, the half circle $\{(x_1, x_2) | x_1^2 + x_2^2 \leq 1 \text{ and } y = -1\}$ there is a fraction ν of individuals with protected attribute equal to -1 who are not treated similarly as individuals with protected value equal to 1.

Sample Complexity The auditing algorithm MMD_{NET} is trained using a decision tree and a unbalanced data ($\mu = 0.2$). Given our setting, for any value of ν we can compute the actual level of differential unfairness γ_o and compare it to the value γ_a estimated by the auditing algorithmMMD_{NET}. Figure 1a plots γ_a against γ_o after 100 simulations for value of β varying from 0 to 0.5 and fixed sub-population size $\alpha = 0.1$. The experiment is conducted with data set of size $n \in \{1000, 5000\}$. The estimated unfairness level γ_a is unbiased since the plots nearly align with 45° line. Larger sample ($5K$) have little effect on how bias the auditor MMD_{NET} is, but reduces the variance of the estimated γ_a.

Unbalanced Data To test the performance of our certifying algorithm on unbalanced data, we repeat the previous experiment with μ varying from -0.3 to 0.3. We compare the performance of three reweighting approaches: uniform weight (UW); direct use of the importance sampling weights



(a) Effect of unfairness intensity β on auditor's performance.



(b) Effect of unbalanced data on auditor's performance.

Figure 1: Certifying γ- multi differential unfairness. The auditor is a decision tree whose depth is tuned using a 5-fold cross validation. The weights function u is obtained using a neural network with four hidden layers with eight neurons each. The plots show the average of 100 experiments. Figure 1a auditor's bias when estimating γ_a is measured by deviations from the 45° line; the unbalance parameter μ is set to 0; the size α of sub-population with a fairness violation is set to 0.1; and, the intensity β of the fairness violation varies from 0 to 0.5. Figure 1b: bias is measured as the difference $\gamma_a - \gamma_o$; the size α of sub-population with a fairness violation is set to 0.1; the intensity β of the fairness violation is set to 0.5; the balancing parameter μ varies from -0.3 to 0.3.

w (ISW); . Figure 1b shows that absent of a reweighting scheme (UW) the certifying algorithm fails to estimate correctly the classifier's unfairness if $\mu \neq 0$. This is in line with the observation made in section 3 that the auditing algorithm needs to control for the information leaked by the auditing features x when measuring the additional leakage from the classifier's outcome y. Our preferred method MMD_{NET} performs well as there is little bias in the estimate γ_a even at large value of the unbalancing factor ν. Note using importance sampling weights directly does worse than a uniform sampling: this confirms previous observations in the literature that in finite sample, the variance of the importance sample weights can affect be detrimental to a re-balancing approach.

Concept Class Figure 4a runs a similar experiment but varies with decision trees of depth varying from 5 to 40. At small level of unfairness, a less complex concept class $\mathbb{C}$ generates less bias in the estimate of γ. A shorter decision tree out-performs significantly deeper structures. This result, in line with the sample complexity presented in theorem 3.2 justifies the concept of multi fairness: in order to be statistically meaningful, the granularity of the sub-populations for which multi differential fairness is audited for is bounded by the complexity of $\mathbb{C}$.

4.2 Case Study: COMPAS

We chose to apply first our method to the COMPAS algorithm because it is a widely used to assess the likelihood of a defendant to become a recidivist. Records from Broward County, Florida are publicly available and expansive analysis of the COMPAS fairness have been already carried out.

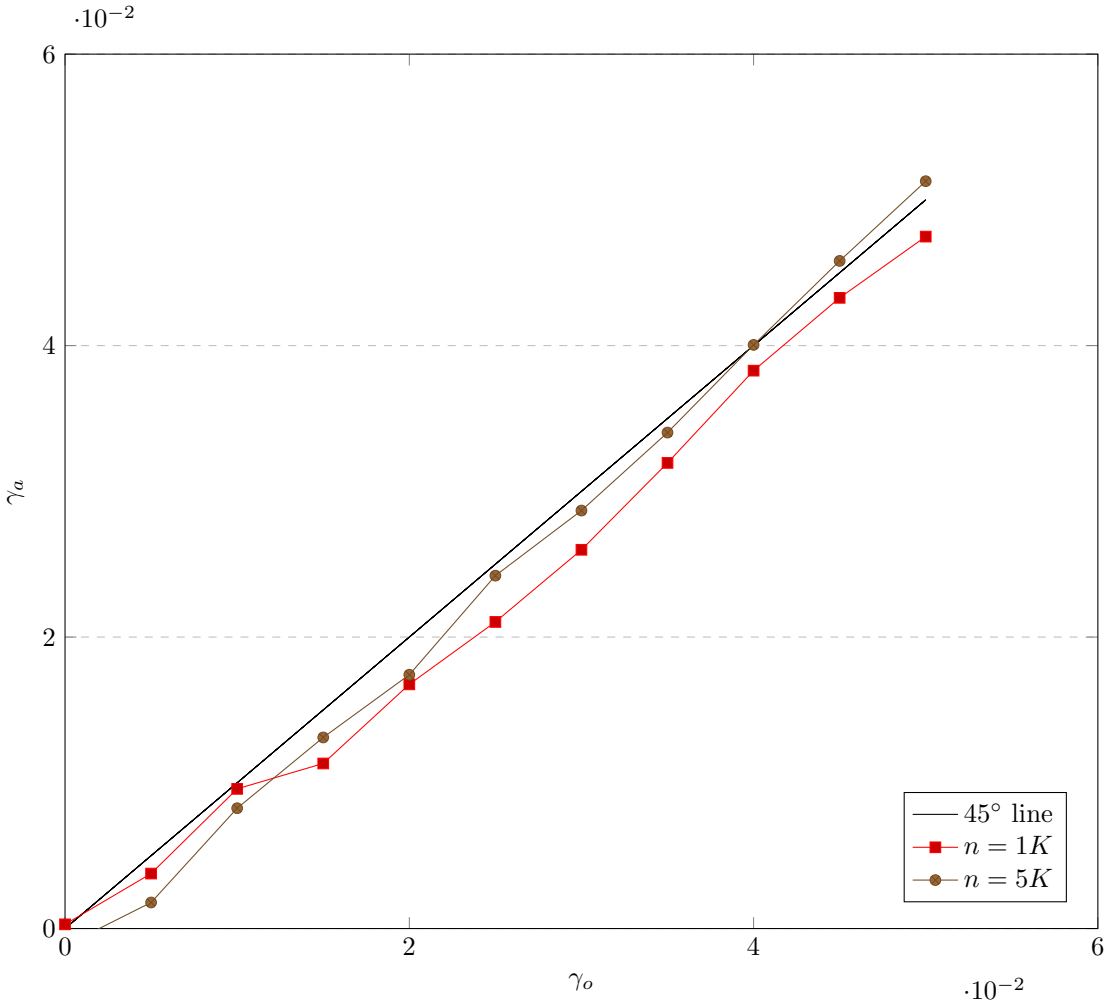


Figure 2: Effect of unfairness intensity β on auditor's performance.

Figure 3: Certifying γ- multi differential unfairness. The auditor is a decision tree whose depth is tuned using a 5-fold cross validation; the unbalance parameter μ is set to 0. The plots show the average of 100 experiments. Auditor's bias when estimating γ_a is measured by deviations from the 45° line. Figure 1a: the size α of sub-population with a fairness violation is set to 0.1; and, the intensity β of the fairness violation varies from 0 to 0.5. Figure 1b: the size α of sub-population with a fairness violation varies from 0.01 to 0.1; and, the intensity β of the fairness violation is set to 0.25.

Our novelty is to apply our multi-differential fairness framework and explore whether there exists in those records group of individuals that could argue for a disparate treatment akin to an "intentional" discrimination.

Data Description The data collected by ProPublica in Broward County from 2013 to 2015 contain $7K$ individuals along with a risk score and a risk category assigned by COMPAS. We transform the risk category into a binary variable equal to 1 for individuals assigned in the high risk category (risk score between 8 and 10). The data provides us with information related to the historical criminal history, misdemeanors, gender, age and race of each individual.

Certifying the Lack of Differential Fairness First, we assess whether the COMPAS risk classification is multi-differential fair. The data is randomly split into train and test sets using 70%/30% ratio. The auditor uses a four layers fully connected neural network to re-balance the data and a random forest to certify differential fairness. The assessment is made for a binary protected attribute: whether an individual self-identified as Afro-American. We find a significant level of differential unfairness in the COMPAS risk classification (see Table **??**) when the binary protected attribute is whether an individual self-identified as Afro-American. The unfairness is slightly stronger with a auditing feature space limited to the count of prior felonies and the degree of the current charge.

Features	Unfairness level (γ)	Standard Deviation
I, II, III, IV, V	0.023761	0.004312
I, II, III	0.023599	0.00384
I, II	0.027382	0.019383

Table 1: Certifying the lack of differential fairness in COMPAS risk classification. Features are as follows: I: count of prior felonies; II: degree of current charge (criminal vs non-criminal); III: age; IV: count of juvenile prior felonies; V: count of juvenile prior misdemeanors. Standard deviations are obtained by drawing 100 train/test splits.

Worst Violations The results in Table **??** do not allow to distinguish the case of a small violation of differential fairness spread evenly across the whole feature distribution from the case of large violations concentrated on small sub-populations.

Disparate Treatment Pick up Alice, Bob, Carol and Dick who belongs to sub-population for which a disparate treatment can be shown.

5 Appendix

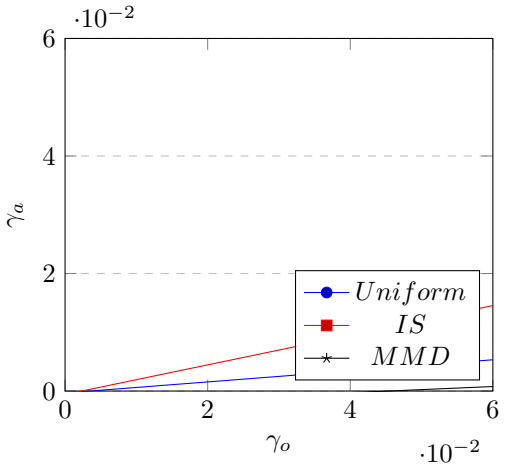
5.1 Analysis of Algorithm 2

Let $a \in \mathfrak{A}$. Without loss of generality we consider the case where $Y = 1$. At each iteration t, denote c_t^* the solution of the following optimization problem

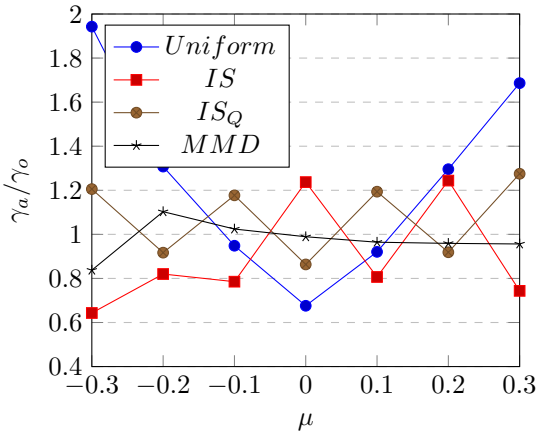
$$max_{c \in \mathbb{C}} E_{D_f^w} \left[\sum_{i=1}^{m} w_{it} \mathbb{1}_a(a_i) c(x_i) \middle| Y = 1 \right], \tag{19}$$

where w_{it} are the weights at iteration t and the expectation is taken over all the samples of size m drawn from D_f^w.

$$w_{it} = \begin{cases} w_i(1 + \nu t) \text{ if } a_i \neq a \\ w_i \text{ otherwise.} \end{cases} \tag{20}$$

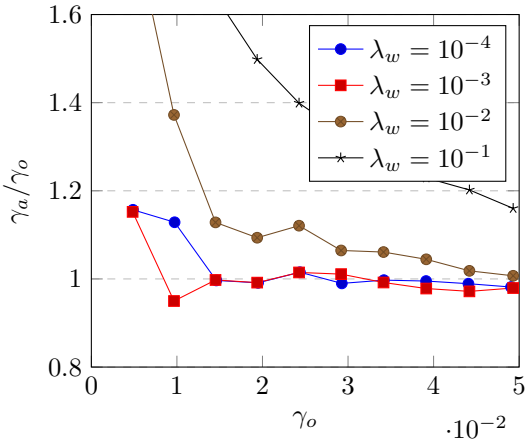


(a) Varying classifier's unfairness γ_o.



(b) Varying data unbalance μ.

Figure 4: Certifying γ- multi differential unfairness when data is unbalanced. Figure 1a: the auditor is a decision tree with depth 5; sample size is $10K$; Figure 1b: the auditor is a decision tree with depth 5; sample size is $10K$; $\gamma_0 = 3$. The auditor's bias is measured by the ratio γ_a/γ_o: a ratio of 1 means that the auditor measures an unbiased level of unfairness.



(a) Varying regularization λ_w.

Figure 5: Certifying γ- multi differential unfairness when data is unbalanced. The auditor is a decision tree with depth 5; sample size is $10K$; The auditor's bias is measured by the ratio γ_a/γ_o: a ratio of 1 means that the auditor measures an unbiased level of unfairness.

For $c \in \mathbb{C}$, denote $\beta_c = E_{c=1,y=1}[\mathbb{1}_a(a_i)]$. Let $c_0 \in \mathbb{C}$ such that $c_0(x) = -1$ for all $x \in \mathfrak{X}$. Denote $B_c = \{x_i | g(x_i) = 1\ f_a(x_i) = f_{a'}(x_i)\}$, $B_g^+ = \{x_i | g(x_i) = 1\ f_a(x_i) > f_{a'}(x_i)\}$ and $B_g^- = \{x_i | g(x_i) =$

$1\ f_a(x_i) < f_{a'}(x_i)\}$. Observe that

$$\beta_g = \left| \sum_{i,x_i \in B_g^+} w_i - \sum_{i,x_i \in B_g^-} w_i \right|. \tag{21}$$

Assume first that $\sum_{i,x_i \in B_g^+} w_i > \sum_{i,x_i \in B_g^-} w_i$. Therefore,

$$\begin{aligned} E_D\left[\sum_{i=1,g(x_i)=1}^{m} w_{it} f_i^* g(x_i)\right] &= E_D\left[\sum_{i=1,x_i \in B_g}^{m} w_{it} f_i^* + \sum_{i=1,x_i \notin B_g}^{m} w_{it} f_i^*\right] \\ &= -\frac{1}{2}\epsilon t E_D[|B_g|] - \epsilon t E_D[|B_g^-|] - E_D\left[\sum_{i=1,x_i \in B_g^-}^{m} w_i\right] \\ &\quad + E_D\left[\sum_{i=1,x_i \in B_g^+}^{m} w_i\right] \\ &= -\frac{1}{2}\epsilon t E_D[|B_g|] - \epsilon t E_D[|B_g^-|] + \beta_g, \end{aligned} \tag{22}$$

where $|B| = \sum_{i=1,x_i \in B}^{m} w_i$ for any $B \in \mathbb{C}$. Moreover,

$$E_D\left[\frac{1}{m}\sum_{i=1,g(x_i)=1}^{m} w_{it} f_i^* g_0(x_i)\right] = \frac{1}{2}\epsilon t E_D[|B_g|] + \epsilon t E_D[|B_g^-|] - \beta_g. \tag{23}$$

Therefore, g cannot be a solution of (19) if

$$\epsilon t \left(E_D[|B_g|] + 2E_D[|B_g^-|]\right) > 2\beta_g. \tag{24}$$

Note that $|B_g| = 1 - \left(|B_g^-| + |B_g^+|\right)$ and $\beta_g = |B_g^+| - |B_g^-|$. Therefore, g cannot be a solution of (19) if

$$t > \frac{2\beta_g}{\epsilon(1-\beta_g)} \tag{25}$$

At any iteration t, a solution of (19) is either $g(x) = -1$ for all $x \in \mathfrak{X}$ or $\beta_g > \frac{\epsilon t}{\epsilon t + 2}$.

Small samples properties We can use a generic uniform convergence property:

Theorem 5.1. *Let$\mathfrak{H}$ be a family of function mapping from $\mathfrak{X}$ to $\{-1, 1\}$ and let $S = \{x_1, ..., x_m\}$ be a sample where $x_i \sim D$ for some distribution D over $\mathfrak{X}$. With probability $1 - \delta$, for all $h \in \mathfrak{H}$*

$$\left| E_{S\sim D}[h] - \frac{1}{m}\sum_{i=1}^{m} h(x_i) \right| \leq 2\mathfrak{R}_m(\mathfrak{H}) + \sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Applying the uniform convergence result from 5.1 allows deriving property of algorithm 2. For any a sample S and any $g \in \mathbb{C}$ with probability $1 - \delta/2$, ,

$$\left| \frac{1}{m} \sum_{i=1, g(x_i)=1}^{m} w_{it} f_i^* g(x_i) + \frac{1}{2}\epsilon t E_D[|B_g|] + \epsilon t E_D[|B_g^-|] - \beta_g \right| \leq 2\mathfrak{R}_m(\mathbb{C}) + \sqrt{\frac{2\ln(2/\delta)}{m}}, \tag{26}$$

and

$$\left| \frac{1}{m} \sum_{i=1, g(x_i)=1}^{m} w_{it} f_i^* g_0(x_i) - \frac{1}{2}\epsilon t E_D[|B_g|] - \epsilon t E_D[|B_g^-|] + \beta_g \right| \leq 2\mathfrak{R}_m(\mathbb{C}) + \sqrt{\frac{2\ln(2/\delta)}{m}}. \tag{27}$$

Therefore, with probability $1 - \delta$, h cannot be solution of the empirical counterpart of (19) if

$$t > \frac{2\beta_g}{\epsilon(1-\beta_g)} + \frac{4}{1-\beta_g}\mathfrak{R}_m(\mathbb{C}) + \frac{2}{1-\beta_g}\sqrt{\frac{2\ln(2/\delta)}{m}}. \tag{28}$$

Therefore at iteration t, with probability $1 - \delta$, a solution of (19) for a sample S is either $g(x) = -1$ for all $x \in S$ or

$$\beta_g > \frac{\epsilon t}{2+\epsilon t} - \frac{4}{2+\epsilon t}\mathfrak{R}_m(\mathbb{C}) - \frac{2}{2+\epsilon t}\sqrt{\frac{2\ln(2/\delta)}{m}}. \tag{29}$$

References

[1] Ricci et al. vs. destefano et al.

[2] How algorithms can bring down minorities credit scores? *The Atlantic*, 2016.

[3] How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

[4] Loomis vs. state of wisconsin. *Supreme Court of the State of Wisconsin*, 2016.

[5] When an algorithm helps send you to prison. *New York Times*, 2017.

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

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[8] 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.

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[11] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

[12] Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2009.

[13] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

[14] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[15] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

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

[17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.