Active Fourier Auditor for Estimating Distributional Properties of ML Models

Ayoub Ajarra
Équipe Scool, Univ. Lille, Inria,
CNRS, Centrale Lille, UMR 9189- CRIStAL
F-59000 Lille, France
ayoub.ajarra@inria.fr

Bishwamittra Ghosh
Max Planck Institute for Software Systems,
Saarbrücken, Germany
bghosh@u.nus.edu

Debabrota Basu
Équipe Scool, Univ. Lille, Inria,
CNRS, Centrale Lille, UMR 9189- CRIStAL
F-59000 Lille, France
debabrota.basu@inria.fr

Abstract

With the pervasive deployment of Machine Learning (ML) models in real-world applications, verifying and auditing properties of ML models have become a central concern. In this work, we focus on three properties: robustness, individual fairness, and group fairness. We discuss two approaches for auditing ML model properties: estimation with and without reconstruction of the target model under audit. Though the first approach is studied in the literature, the second approach remains unexplored. For this purpose, we develop a new framework that quantifies different properties in terms of the Fourier coefficients of the ML model under audit but does not parametrically reconstruct it. We propose the Active Fourier Auditor (AFA), which queries sample points according to the Fourier coefficients of the ML model, and further estimates the properties. We derive high probability error bounds on AFA's estimates, along with the worst-case lower bounds on the sample complexity to audit them. Numerically we demonstrate on multiple datasets and models that AFA is more accurate and sample-efficient to estimate the properties of interest than the baselines.

1 Introduction

As Machine Learning (ML) systems are pervasively being deployed in high-stake applications, mitigating discrimination and guaranteeing reliability are critical to ensure the safe pre and post-deployment of ML [Madiega, 2021]. These issues are addressed in the growing subfield of ML, i.e. trustworthy or responsible ML [Rasheed et al., 2022, Li et al., 2023], in terms of robustness and fairness of ML models. Robustness quantifies how stable a model's predictions are under perturbation of its inputs [Xu and Shie, 2011, Kumar et al., 2020]. Fairness [Dwork et al., 2012, Barocas et al., 2023] seeks to address discrimination in predictions both at the individual level and across groups. Thus, AI regulations, such as the European Union AI Act [Madiega, 2021], increasingly suggest certifying different model properties, such as robustness, fairness, and privacy, for a safe integration of of ML in high-risk applications. Thus, estimating these model properties under minimum interactions with the models has become a central question in algorithmic auditing [Raji et al., 2020, Wilson et al., 2021, Metaxa et al., 2021, Yan and Zhang, 2022].

2nd Workshop on Regulatable ML at NeurIPS 2024.

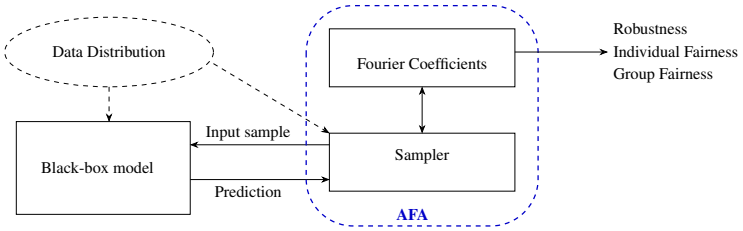


Figure 1: A schematic of AFA.

Example 1. Following [Ghosh et al., 2021, Example 1], let us consider an ML model that predicts who is eligible to get medical insurance given a sensitive feature 'age', and two non-sensitive features 'income' and 'health'. Owing to historical bias in the training data, the model, i.e. an explainable decision tree, discriminates against the 'elderly' population by denying their health insurance and favors the 'young' population. Hence, an auditor would realize that the model does not satisfy *group fairness* since the difference in the probability of approving health insurance between the elderly and the young is large. In addition, the model violates *individual fairness*, where perturbing the feature 'age' from elderly to young increases the probability of insurance. Further, the model lacks *robustness* if perturbing any feature by an infinitesimal quantity flips the prediction.

Related Work: ML Auditing. Towards trustworthy ML, several methods have been proposed to ally audit an ML model by estimating different *distributional properties* of it, such as fairness and robustness, where the model hyper-property has to be assessed against the distribution of inputs. A stream of work focuses on property verification that verifies whether these properties are violated above a pre-determined threshold [Goldwasser et al., 2021, John et al., 2020, Mutreja and Shafer, 2023, Herman and Rothblum, 2022, Kearns et al., 2018]. Thus, we focus on estimating these properties instead of a 'yes/no' answer, which is a harder problem than verification [Goldwasser et al., 2021]. On estimating distributional properties, Neiswanger et al. [2021] proposed a Bayesian approach for estimating properties of black-box optimizers and required a prior distribution of models. Wang et al. [2022] studies simpler distributional properties, e.g. the mean, the median, and the trimmed mean defined as a conditional expectation, using offline and interactive algorithms. Yan and Zhang [2022] considered a frequentist approach for estimating group fairness but assumed the knowledge of the model class and a finite hypothesis class under audit. These assumptions are violated if we do not know the model type and can be challenging for complex models, e.g. deep neural networks. Albarghouthi et al. [2017], Ghosh et al. [2021] considered finite models for estimating group fairness w.r.t. the features distribution, and Ghosh et al. [2022] further narrowed down to linear models. Therefore, we identify the following limitations of the existing methods in ML auditing. (1) **Property-specific auditing:** most methods considered a property-specific tailored approach to audit ML systems, for example either robustness [Cohen et al., 2019, Salman et al., 2019], group fairness [Albarghouthi et al., 2017, Ghosh et al., 2021], or individual fairness [John et al., 2020]. (2) **Model-specific auditing:** all the methods considered a prior knowledge about the ML model [Neiswanger et al., 2021, Ghosh et al., 2021, 2022, Yan and Zhang, 2022], or a white-box access to it [Cohen et al., 2019, Salman et al., 2019]. These are unavailable in practical systems such as API-based ML. Therefore, our research question is: *Can we design a unified ML auditor for black-box systems for estimating a set of distributional properties including robustness and fairness?*

Contributions. We propose a framework, namely AFA (**A**ctive **F**ourier **A**uditor), which is an ML auditor based on the Fourier approximation of a black-box ML model (Figure 1). We observe that existing black-box ML auditors work in two steps: *the model reconstruction step*, where they reconstruct a model completely, and *the estimation step*, where they put an estimator on top of it [Yan and Zhang, 2022]. We propose a model-agnostic strategy that does not need to reconstruct the model completely. In particular, for any ML model admitting a Fourier expansion, we compute the significant Fourier coefficients of a model accepting categorical input distributions such that they are enough to estimate different distributional properties such as robustness, individual fairness, and group fairness. Our contributions are:

- *Formalism.* For any bounded output model (e.g. all classifiers), we theoretically reduce the estimation of robustness, individual fairness, and group fairness in terms of the Fourier coefficients of the model. The key idea is based on influence functions, which capture how much a model output changes due to a change in input variables and can be computed via

Fourier coefficients (Section 3). We propose two types of influence functions for each of these properties that unifies robustness and individual fairness auditing while put group fairness in a distinct class.

- *Algorithm.* In AFA, we integrate Goldreich-Levin algorithm [Goldreich and Levin, 1989, Kushilevitz and Mansour, 1993] to efficiently compute the significant Fourier coefficients of the ML model, which are enough to compute the corresponding properties. AFA yields a probably approximately correct (PAC) estimation of distributional properties. We propose a dynamic version of Goldreich-Levin to accelerate the computations.
- *Theoretical Sample Complexity.* We show that our algorithm requires $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\sqrt{\log\frac{1}{\delta}}\right)$ samples to yield (ϵ, δ) estimate of robustness and individual fairness, while it needs $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples to audit group fairness. We further derive a lower bound on the sample complexity of (ϵ, δ)-auditing of group fairness to be $\tilde{\Omega}(\frac{\delta}{\epsilon^2})$. Further, for group fairness, we prove that AFA is manipulation-proof under perturbation of 2^{n-1} Fourier coefficients.
- *Experimental Results.* We numerically test the performance of AFA to estimate the three properties of different types of models. The results show that AFA achieves lower estimation error while estimating robustness and individual fairness across perturbation levels. Compared to existing group fairness auditors, AFA not only achieves lower estimation error but also incurs lower computation time across models and the number of samples.

2 Background

Before proceeding to the contributions, we discuss the three statistical properties of ML models that we study, i.e. robustness, individual fairness, and group fairness. We also discuss basics of Fourier analysis that we leverage to design AFA.

Notations. Here, x represents a scalar, and $\mathbf{x}$ represents a vector. $\mathcal{X}$ is a set. We denote $[\![1, n]\!]$ as the set $\{1, \ldots, n\}$. We denote the power set of $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$.

Properties of ML Models. A Machine Learning (ML) model h is a deterministic or probabilistic mapping from an n-dimensional input domain of features (or covariates) $\mathcal{X}$ to set of labels (or response variables or outcomes) $\mathcal{Y}$. For example, for Boolean features $\mathcal{X} \triangleq \{-1, 1\}^n$, and for categorical features, $\mathcal{X} \triangleq [K]^n$. For binary classifiers, $\mathcal{Y} \triangleq \{0, 1\}$.

We assume to have only *black-box access to* h, i.e. we send queries from a data-generating distribution and collect only the labels predicted by h. The dataset on which h is tested is sampled from a data-generating distribution $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$ over $\mathcal{X} \times \mathcal{Y}$, which has a marginal distribution $\mathcal{D}$ over $\mathcal{X}$.

We aim to audit a distributional *(aka global)* property $\mu : \mathcal{H} \times \mathcal{D}_{\mathcal{X},\mathcal{Y}} \to \mathbb{R}$ *of an ML model* $h : \mathcal{X} \to \mathcal{Y}$ *belonging to an unknown model class* $\mathcal{H}$ *while having only black-box access to* h.

Hereafter, *we develop the methodology for binary classifiers and Boolean features*. Later, we discuss approaches to extend the proposed methodology to categorical features and multi-class classifiers, and corresponding experimental results. In this paper, we study three properties of ML models, i.e. robustness (μ_{Rob}), individual fairness (μ_{IFair}), and group fairness (μ_{GFair}), which are defined below.

Robustness is the ability of a model h to generate same output against a given input and its perturbed (or noisy) version. Robustness has been central to sub-fields of AI, e.g. safe RL [Garcıa and Fernández, 2015], adversarial ML [Kurakin et al., 2016, Biggio and Roli, 2018], and gained attention for safety-critical deployment of AI.

Definition 1 (Robustness)**.** *Given a model* h *and a perturbation mechanism* Γ *of input* $x \in \mathcal{X}$*, robustness of* h *is* $\mu_{\mathrm{Rob}}(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\ \mathbf{y}\sim\Gamma(x)}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

Examples of perturbation mechanisms include Binary feature flipping $N_\rho(\mathbf{x}) \triangleq \{\mathbf{x}' \mid \forall i \in [n], \mathbf{x}'_i = \mathbf{x}_i \times \mathrm{Bernoulli}(\rho)\}$ [O'Donnell, 2014], Gaussian perturbation $N_\rho(x) \triangleq \{\mathbf{x}' \mid \mathbf{x}' = \mathbf{x} + \epsilon \text{ where } \epsilon \sim \mathrm{Normal}(0, \rho^2 I)\}$ [Cohen et al., 2019], among others.

In trustworthy and responsible AI, another prevalent concern about deploying ML models is bias in their predictions. This has led to the study of different fairness metrics, their auditing algorithms, and algorithms to enhance fairness [Mehrabi et al., 2021, Barocas et al., 2023]. There are two categories

Table 1: Example 3

S	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
χ_S	1	x_1	x_2	x_1x_2	x_3	x_1x_3	x_2x_3	$x_1x_2x_3$
ψ_S	1	x_1	x_2	x_1x_2	0	0	0	0

of fairness measures [Barocas et al., 2023]. The first is the **individual fairness** that aims to ensure that individuals with similar features should obtain similar predictions [Dwork et al., 2012].

Definition 2 (Individual Fairness)**.** *For a model h and a neighbourhood $\Gamma(x)$ of a $x \in \mathcal{X}$, the individual fairness discrepancy of h is $\mu_{\mathrm{IFair}}(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\, \mathbf{y}\sim\Gamma(\mathbf{x})}\, \mathbb{P}[h(\mathbf{x}) \neq h(\mathbf{y})]$.*

The neighborhood $\Gamma(x)$ is commonly defined as the points around x which are at a distance less than $\rho \geq 0$ w.r.t. a pre-defined metric. The metric depends on the application of choice and the input data [Mehrabi et al., 2021]. IF of a model measures its capacity to yield similar predictions for similar input features of individuals [Dwork et al., 2012, A. Friedler et al., 2016]. The similarity between individuals are measured with different metrics. Let $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ be the metrics for the metric spaces of input ($\mathcal{X}$) and predictions ($\mathcal{Y}$), respectively.

A model h satisfies (ϵ, ϵ')-IF if $d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \leq \epsilon$ implies $d_{\mathcal{Y}}(h(\mathbf{x}), h(\mathbf{x}')) \leq \epsilon'$ for all $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$ [A. Friedler et al., 2016]. For Boolean features and binary classifiers, the natural candidate for $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ is the *Hamming distance*. This measures the difference between vectors $\mathbf{x}$ and $\mathbf{x}'$ by counting the number of differing elements. Thus, $d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \leq l$ means that $\mathbf{x}'$ has l different bits than $\mathbf{x}$. As auditors, we are interested in measuring how much the Hamming distance between outcomes of $\mathbf{x}$ and $\mathbf{x}'$, i.e. ϵ'. However, since the data-generation process and the models might be stochastic, we take a stochastic view and use a perturbation mechanism that defines a neighborhood around each input sample.

Group fairness is the other category of fairness measures that considers the input to be generated from multiple protected groups (or sub-populations), and we want to remove discrimination in predictions across these protected groups [Mehrabi et al., 2021]. Specifically, we focus on *Statistical Parity (SP)* [Feldman et al., 2015, Dwork et al., 2012] as our measure of deviation from group fairness. For simplicity, we discuss SP for two groups, but we can also generalize it to multiple groups.

Definition 3 (Statistical Parity)**.** *The statistical parity of h is $\mu_{\mathrm{GFair}}(h) \triangleq |\, \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = 1] - \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = -1]|$, where x_A is the binary sensitive attribute.*

In AFA*, we use techniques of Fourier analysis to design one computational scheme for simultaneously estimating these three properties of an ML model.*

A Primer on Fourier Analysis. Designing AFA is motivated by the Fourier expansion of Boolean functions. Fourier coefficients are distribution-dependent components that capture key information about the distribution's properties. This study was initially addressed by [O'Donnell, 2014], who focused on the uniform distribution. Later, [Heidari et al., 2021] generalized this result to arbitrary distributions, which we leverage further.

Proposition 1 (Heidari et al. [2021])**.** *There exists a set of orthonormal parity functions $\{\psi_S\}_{S\subseteq[n]}$ such that any function $h : \{-1, 1\}^n \to \{-1, 1\}$ is decomposed as*

$$h(x) = \sum\nolimits_{S\subseteq[n]} \hat{h}(S)\psi_S(\mathbf{x}) \;\; \textit{for any} \;\; x \sim \mathcal{D}. \tag{1}$$

The Fourier coefficients $\hat{h}(S) \triangleq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[h(X)\psi_S(\mathbf{x})]$ are unique for all $S \subseteq [n]$.

Example 2. *Let us consider h to be the XOR function on $\mathbf{x} \in \{-1, 1\}^2$. This means that $h(-1, -1) = h(1, 1) = 0$ and $h(1, -1) = h(-1, 1) = 1$. The Fourier representation of $h(\mathbf{x}) = 0.5 + 0.5x_1 + 0.5x_2 - 0.5x_1x_2$, when $\mathbf{x}$ is sampled from a uniform distribution on $\{-1, 1\}^2$.*

Example 3. *Suppose random variables X_1 and X_2 are drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, 1)$ [Heidari et al., 2021]. Define another random variable X_3 as $X_3 = X_1X_2$. It can be verified that the Gram-Schmidt basis of XOR of X_1, X_2, X_3 has four zero coefficients, i.e. the sets including X_3 do not influence the outcomes. This is because X_3's information is encoded in X_1 and X_2 jointly.*

Influence functions. To estimate the properties of interest, we use a tool from Fourier analysis, i.e. *influence functions* [O'Donnell, 2014]. They measure how changing an input changes the output of a model. Different influence functions are widely used in statistics, e.g. to design robust estimators [Mathieu et al., 2022], and ML, e.g. to find important features [Heidari et al., 2021], to evaluate how features induce bias [Ghosh et al., 2021], to explain contribution of datapoints on predictions [Ilyas et al., 2022]. Here, we use them to estimate model properties.

Definition 4 (Influence functions)**.** *If Γ is a transformation of an input $\mathbf{x} \in \mathcal{X}$, the influence function is defined as $\mathrm{Inf}_\Gamma(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \neq h(\Gamma(\mathbf{x}))]$. $\mathrm{Inf}_\Gamma(h)$ is called deterministic if the transformation Γ is deterministic, and randomized if Γ randomized.*

In general, deterministic influence functions are used in Boolean function analysis [O'Donnell, 2014]. In contrast, in Section 3, we express robustness, individual fairness, and group fairness with randomized influence functions. We also show that the influence functions can be computed using the Fourier coefficients of the model under audit (Equation (1)).

3 Active Fourier Auditor

In the black-box setting, the access to the model h is limited by the query oracle, accessible to the auditor. The auditor's objective is to estimate the property μ through interaction with this oracle. The definition of the property estimator relies on the information made available to the auditor during this interaction. In the context of auditing with model reconstruction [Yan and Zhang, 2022], the auditor is denoted as $\hat{\mu} : \mathcal{H} \times \mathcal{B} \to \mathbb{R}$. Here, the auditor has access to an unlabeled pool and applies active learning techniques (e.g. CAL algorithm) to query samples. This process uses the additional information given by the hypothesis class where the model h lives. Following the reconstruction phase, the auditor has an approximate model $\hat{h}$ of true model h, enabling estimation of the property via plug-in estimator $\hat{\mu}(\hat{h})$.

Now, we present a novel non-parametric black-box auditor AFA that assumes no knowledge of the model class and the data-generating distribution. Unlike the full model-reconstruction-based auditors, AFA uses Fourier expansion and adaptive queries to estimate the robustness, Individual Fairness (IF), and Group Fairness (GF) properties of a model h. In this setting, the auditor is defined as $\hat{\mu} : \mathcal{F}_\mu \times \mathcal{B} \to \mathbb{R}$, where $\mathcal{F}_\mu$ represents the set of Fourier coefficients upon which the property μ depends. First, we show that property estimation with model reconstruction always incurs higher error. Then, we show that robustness, IF, and GF for binary classifiers can be computed using Fourier coefficients of h. Finally, we compute the Fourier coefficients and thus, estimate the properties at once (Algorithm 1). We begin by defining a PAC-agnostic auditor that we realise with AFA.

Definition 5 (PAC-agnostic auditor)**.** *Let μ be a computable distributional property of model h. An algorithm $\mathcal{A}$ is a PAC-agnostic auditor if for any $\epsilon, \delta \in (0,1)$, there exists a function $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$ samples drawn from $\mathcal{D}$, it outputs an estimate $\hat{\mu}_m$ satisfying $\mathbb{P}(|\hat{\mu}_m - \mu| \leq \epsilon) \geq 1 - \delta$.*

Remark. $\mu(h)$ is a *computable* property if there exists a (randomized) algorithm, such that when given access to (black-box) queries, it outputs a PAC estimate of the property $\mu(h)$ [Kearns et al., 2018]. Any distributional property, including robustness, individual fairness and group fairness, is computable given the existence of the uniform estimator.

3.1 The Cost of Reconstruction

The naive way to estimate a model property is to reconstruct the model and then use a plug-in estimator [Yan and Zhang, 2022]. However, this requires an exact knowledge of the model class and comes with an additional cost of reconstructing the model before property estimation. For group fairness, we show that the reconstruct-then-estimate approach induces significantly higher error than the reconstruction error, while the exact model reconstruction itself is NP-hard [Jagielski et al., 2020].

Proposition 2. *If $\hat{h}$ is the reconstructed model from h, then*

$$|\mu_{\mathrm{GFair}}(\hat{h}) - \mu_{\mathrm{GFair}}(h)| \leq \min\left\{1, \frac{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})]}{\min(\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = 1], \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = -1])}\right\}.$$

Proposition 2 connects the estimation error and the reconstruction error before plugging in the estimator. It also shows that to have a sensible estimation the reconstruction algorithm needs to

achieve an error below the proportion of minority group, which can be significantly small requiring high sample complexity. The proof is deferred to Appendix A. This motivates an approach that avoids model reconstruction by computing only the right components of the model expansion. To capture the information relevant to estimating our properties of interest, we will represent them in terms of Fourier coefficients given in the model decomposition. Then we aim to adaptively estimate larger Fourier coefficients in contrast to model reconstruction method requiring to recovering all the Fourier coefficients.

3.2 Model Properties with Fourier Expansion

Throughout the rest of this paper, we denote by $\{\psi_S\}_{S\subseteq[n]}$ the basis derived from Proposition 1. In this section, we express the model properties of h using its Fourier coefficients. The detailed proofs are deferred to Appendix B.

a. Robustness. Robustness of a model h measures its ability to maintain its performance when new data is corrupted. Auditing robustness requires a generative model to imitate the corruptions, which is modelled by the perturbation mechanism (Definition 1). As we focus on the Boolean case, the worst case perturbation Γ_ρ is the protocol of flipping vector coordinates with a probability ρ. Specifically, a corrupted sample $\mathbf{y}$ is generated from $\mathbf{x}$ such that for every component, we independently set $y_i = x_i$ with probability $\frac{1+\rho}{2}$ and $y_i = -x_i$ with probability $\frac{1-\rho}{2}$. This perturbation mechanism leads us to the ρ-flipping influence function.

Definition 6 (ρ-flipping Influence Function)**.** *The ρ-flipping influence function of any model h is defined as* $\mathrm{Inf}_\rho(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

For a Boolean classifier, we further observe that $\mathrm{Inf}_\rho(h) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\ \mathbf{y}\sim N_\rho(x)}[h(\mathbf{x})h(\mathbf{y})]$. This allows us to show that the robustness of h under Γ_ρ perturbation is measured by ρ-flipping influence function, and thus, can be computed using Fourier coefficients of h.

Proposition 3. *Robustness of h under the Γ_ρ flipping perturbation is equivalent to the ρ-flipping influence function, and thus, can be expressed as*

$$\mu_{\mathrm{Rob}}(h) = \mathrm{Inf}_\rho(h) = \sum_{S\subseteq[n]} \rho^{|S|}\hat{h}(S)^2. \tag{2}$$

b. Individual Fairness (IF). To demonstrate the universality of our approach, we express IF with the model's Fourier coefficients. We consider the perturbation mechanism $\Gamma = \Gamma_{\rho,l}(\cdot)$ that independently flips uniformly l vector coordinates with a probability $\frac{1+\rho}{2}$. Thus, we consider a neighbourhood with $\mathbb{E}_{\mathbf{x}'\sim\Gamma_{\rho,l}(\mathbf{x})}[d_{\mathcal{X}}(\mathbf{x},\mathbf{x}')] \leq \frac{1}{2}(1+\rho)l$ around each sample $\mathbf{x}$ as the similar set of individuals. This perturbation mechanism leads us to the (ρ, l)-flipping influence function.

Definition 7 ((ρ, l)-flipping influence function)**.** *The (ρ, l)-flipping influence function of any model h is defined as* $\mathrm{Inf}_{\rho,l}(h) = \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim N_{\rho,l}(\mathbf{x})}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

We leverage (ρ, l)-flipping influence function to express IF of h in terms of its Fourier coefficients (Proposition 4).

Proposition 4. *Individual fairness defined with respect to the $\Gamma_{\rho,l}$ perturbation is equivalent to the (ρ, l)-flipping influence function, and thus, can be expressed as*

$$\mu_{\mathrm{IFair}}(h) = \mathrm{Inf}_{\rho,l}(h) = \sum_{S\subseteq[n]} \rho^{|S_l|}\hat{h}(S)^2\,, \tag{3}$$

where S_l denotes the power sets for which l features change.

Unifying robustness and IF: The Characteristic Function. It is worth noting that IF is similar to robustness, differing only by a single degree of freedom, i.e. the number of flipped directions l. Specifically, from Equation (2) and (3), we observe that both the properties as $\mu(h) = \sum_{S\subseteq[n]} \texttt{char}(S, \mu_\cdot)\hat{h}(S)^2$, such that $\texttt{char}(S, \mu_{\mathrm{Rob}}) = \rho^{|S|}$, and $\texttt{char}(S, \mu_{\mathrm{IFair}}) = \rho^{|S_l|}$. We call `char` as *the characteristic function of the property*.

c. Group Fairness (GF). Now, we focus on Group Fairness which aims to ensure similar predictions for different subgroups of population [Barocas et al., 2023]. We focus on Statistical Parity (SP)

as the measure of deviation from GF [Feldman et al., 2015]. To quantify SP, we propose a novel membership influence function.

Definition 8 (Membership influence function)**.** *If A denotes a sensitive feature, we define the membership influence function w.r.t. A as the conditional probability* $\mathrm{Inf}_A(h) \triangleq \mathbb{P}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}}\Big[h(\mathbf{x}) \neq h(\mathbf{y})\Big| x_A = 1, y_A = -1\Big]$.

$\mathrm{Inf}_A(h)$ is the conditional probability of the change in the outcome of h due to change in group membership of samples from $\mathcal{D}$. In other words, it expresses the amount of independence between the outcome and group membership.

Note that the membership influence function is a randomised version of the deterministic influence function in [O'Donnell, 2014]. If we denote the transformation of flipping membership, i.e. sensitive attribute of $\mathbf{x}$, $f_A(\mathbf{x})$, the classical influence function is $\mathrm{Inf}_A^{det} = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \neq h(f_A(\mathbf{x}))]$. The limitation of this deterministic function is that given $\mathbf{x} \sim \mathcal{D}$ the transformed vector $f_A(\mathbf{x})$ may not represent a sample from $\mathcal{D}$. Thus, it fails to encode the information relevant to SP, whereas the proposed membership influence function does it correctly as shown below.

Proposition 5. *Statistical parity of h w.r.t a sensitive attribute A and distribution $\mathcal{D}$ is the root of the second order polynomial $P_{\hat{h}}(X)$, i.e. $\alpha(1-\alpha)X^2 - \hat{h}(\emptyset)(1-2\alpha)X - \sum_{S\subseteq[n], S\ni A}\hat{h}(S)^2 - \frac{(1-\hat{h}^2(\emptyset))}{2}$, where $\alpha = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[x_A = 1]$ and $\hat{h}(\emptyset)$ is the coefficient of empty set.*

Summary of the Fourier Representation of Model Properties. Robustness and individual fairness have the same Fourier pattern. They depend on all the Fourier coefficients of the model but differ only on their characteristic functions. In contrast, statistical parity of a sensitive feature A depends only on the Fourier coefficient of that sensitive feature $\hat{h}(\{A\})$ and the Fourier coefficient of the empty set $\hat{h}(\emptyset)$.

3.3 NP-hardness of Exact Computation

We have shown that the exact computation of robustness and individual fairness depends on all Fourier coefficients of the model. Since each Fourier coefficient of h is given by $\hat{h}(S) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\psi_S(\mathbf{x})]$, exactly computing a single Fourier coefficient takes $\mathcal{O}(|\mathcal{X}|)$ time. Additionally, the number of Fourier coefficients to compute to estimate robustness and individual fairness is exponential in the dimension of the input domain (2^n). Thus, exactly computing robustness and individual fairness requires $\mathcal{O}(2^n|\mathcal{X}|)$ time. This gives us an idea about the computational hardness of the exact estimation problem. Now, we prove estimating large Fourier coefficients to be NP-complete.

Theorem 1. *Let $\mathcal{Q} \triangleq \{\mathbf{x}, h(\mathbf{x})\}$ be the set of input samples sent to h and the predictions obtained. Given $\tau \in \mathbb{R}_{\geq 0}$, exactly computing all the τ-significant Fourier coefficients of h is NP-complete.*

Proof Sketch. For a set of queries $\mathcal{Q}$ and for each power set S, Fourier coefficient is given by $\hat{h}(S) = \frac{1}{|\mathcal{Q}|}\sum_{(x,h(x))\in\mathcal{Q}} h(x)\psi_S(x)$. Maximizing the Fourier coefficient $|\hat{h}(S)|$ is equivalent to maximizing the agreement or disagreement between h and the sign of ψ_S for each truth assignment. Alternatively, maximizing $|\hat{h}(S)|$ is equivalent to finding a truth assignment that maximizes the number of true clauses in a CNF, where each clause is a disjunction of $h(x)$ and the sign of $\psi_S(x)$, and the CNF includes all such clauses for all $x \in \mathcal{Q}$. This is known as the Max2Sat (maximum two satisfiability) problem, which is known to be NP-complete. Hence, we conclude that finding large Fourier coefficients is also NP-complete. This result shows that the exact computation of the Fourier coefficients for our properties is NP-hard. This has motivated us to design AFA, which we later proved to be an (ϵ, δ)-PAC agnostic auditor.

3.4 Algorithm: Active Fourier Auditor (AFA)

We have shown that finding significant Fourier coefficients can be an NP-hard problem. In this section, we propose AFA (Algorithm 1) that takes as input a *restricted access* of $q > 0$ queries from the data-generating distribution and requests labels from the black-box oracle of h (Line 2). Those queries enable us to find the squares of significant Fourier coefficients and estimate them

Algorithm 1 Active Fourier Auditor (AFA)

1: **Input**: Sensitive attribute A, Query access to h, $\tau, \delta \in (0,1)$, $\epsilon \leftarrow \tau^2/4$
2: $\{x_k, h(x_k)\}_{k\in[q]} \leftarrow$ BLACKBOXQUERY(h, q)
3: $L_h \leftarrow$ GOLDREICHLEVIN(h, q, τ, δ)
4: $\hat{\mu}(h) \leftarrow \sum_{S\in L_h} \texttt{char}(\mu, S)\hat{h}(s)^2$
5: $\hat{\mu}_{GF}(h) \leftarrow P_{\hat{h}}^{-1}(0)$
6: **return** $\{\hat{\mu}_{RB}, \hat{\mu}_{IF}, \hat{\mu}_{GF}\}$

simultaneously. The list of the significant Fourier coefficients L_h of the model h contains both subsets and their estimated Fourier weights. We adopt a Goldreich-Levin (GL) algorithm based approach [Goldreich and Levin, 1989, Kushilevitz and Mansour, 1993] to find such list of significant Fourier coefficients (Figure 2). Since estimating the properties – robustness, individual fairness and group fairness – depend on estimating those Fourier coefficients, we plug in their computed estimates and output an (ϵ, δ)-PAC estimate of the properties (Line 4 and 5).

Algorithmic Insights. To compute the significant Fourier coefficients, we start with the power set. Now, we denote the subsets containing an element i as $\mathcal{B}_i(\mathcal{X})$, and the subsets not containing i as $\mathcal{B}_{\neg i}(\mathcal{X})$. Let Υ denote a trajectory starting from the set of all Fourier coefficients in the binary search tree of Fourier coefficients (Figure 1). The question is that *from the power set, how can we design a Υ to reach subsets of Fourier coefficients above a given threshold τ?*

In AFA, we dynamically create "buckets" of coefficients for this purpose. Each bucket $\mathcal{B}^{S,k}$, represents a collection of power sets, such that $\mathcal{B}^{S,k} \triangleq \{S \cup T \mid T \subseteq \{k+1, \ldots, n\}\}$. The corresponding weight is quantified by $\mathcal{W}^{S,k} \triangleq \sum_{T\subseteq\{k+1,\ldots,n\}} \hat{h}(S \cup T)^2$. In this context, $\mathcal{W}^{S,k}$ measures the total contribution of the Fourier coefficients associated with the elements in the bucket $\mathcal{B}^{S,k}$. The bucket is initialized at $\mathcal{B}^{\emptyset,0}$, which represents the weight of the power set of $[\![1, n]\!]$. By Parseval's identity, we know that the weight of the power set is 1, i.e. $\sum_{S\in\mathcal{P}(\mathcal{X})} \hat{h}(S)^2 = 1$. The bucket $\mathcal{B}^{S,k}$ is then split into two buckets of the same cardinal: $\mathcal{B}^{S,k-1}$ and $\mathcal{B}^{S\cup\{k+1\},k+1}$. We then estimate the weight of each bucket by sending black-box queries to the model h. The algorithm discards the bucket whose weight is below the threshold. When all the buckets collected at a round consist of exactly one element each, i.e. we reach the leaves, the algorithm halts and the buckets collected in this process are subsets of $[\![1, n]\!]$ that have large Fourier coefficients.

Extension to Continuous Features. Heidari et al. [2021] extend Proposition 1 to encompass a general Euclidean space. We use the generic construction of Fourier coefficients in the Euclidean space to extend our computations for feature spaces involving both categorical and continuous features. Rest of our computations follow naturally.

Extension to Multi-class Classification. We also deploy AFA for multi-class classification, where $\mathcal{Y}$ consists of multiple labels. In this setting, the concept of group fairness, i.e. $\mu_{\text{GFair}}(h) \triangleq \max_{y\in\mathcal{Y}} |\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = y|x_A = 1] - \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = y|x_A = -1]|$, is called multicalibration [Dwork et al., 2023]. Here, we construct Fourier expansions of the model for each pair of labels. Then, we use Proposition 5 to compute the group fairness for each of the expansions, and finally, take the maximum to estimate multi-group fairness of h. Formal details are deferred to Appendix D. We experimentally evaluate both the extensions.

4 Theoretical Analysis

Upper Bounds on Sample Complexity.

Theorem 2 (Upper bounds for Robustness and Individual Fairness)**.** AFA *is a PAC-agnostic auditor for robustness and individual fairness with sample complexity* $\mathcal{O}\Big(\frac{\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\Big)$. *Here,* $\texttt{char}(L, \mu_\cdot) \triangleq \sum_{S\in L} \texttt{char}(S, \mu_\cdot)$ *and* $\texttt{char}(\bar{L}, \mu_\cdot) \triangleq \sum_{S\in\bar{L}} \texttt{char}(S, \mu_\cdot)$.

Theorem 3 (Upper bounds for Group Fairness)**.** AFA *yields an (ϵ, δ)-PAC estimate of $\mu_{\text{GFair}}(h)$ if it has access to predictions of $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{4}{\delta}\right)$ input samples.*

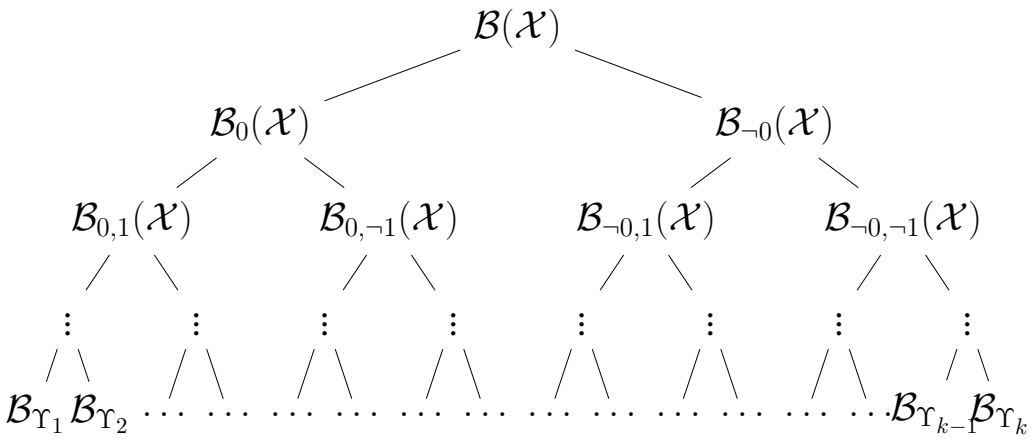


Figure 2: AFA begins with the set of all Fourier coefficients, with weight 1, which is above the threshold $\tau < 1$. It proceeds by splitting the bucket and verifies at each level of the tree the weight of the node. If the weight is below the threshold, the algorithm halts. Otherwise, it continues to expand, yielding a set of (informative) trajectories Υ, the subsets with large Fourier coefficients are $\{\mathcal{B}_{\Upsilon_1}(\mathcal{X}), \cdots, \mathcal{B}_{\Upsilon_k}(\mathcal{X})\}$.

We prove that AFA achieves an optimal rate of $\tilde{\mathcal{O}}(\frac{1}{\epsilon}\sqrt{\log \frac{1}{\delta}})$ for robustness and individual fairness and an $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ rate for group fairness. Consequently, under the same number of samples, AFA exhibits a higher error rate for group fairness compared to robustness and individual fairness, as group fairness involves solving a quadratic equation while the others correspond to their respective influence functions. The proofs of these theorems are in Appendix C.

Corollary 1. AFA *yields an* (ϵ, δ)*-PAC estimate of robustness, individual fairness, and group fairness in time complexity poly* $\left(\frac{1}{\epsilon}, n\right)$ *using:*

- $\mathcal{O}\Big(\frac{\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log \frac{2}{\delta}}\Big)$ *queries for robustness and individual fairness.*
- $\mathcal{O}\left(\frac{1}{\epsilon^2} \log \frac{4}{\delta}\right)$ *queries for group fairness.*

Where, $\texttt{char}(L, \mu.) \triangleq \sum_{S \in L} \texttt{char}(S, \mu.)$ *and* $\texttt{char}(\bar{L}, \mu.) \triangleq \sum_{S \in \bar{L}} \texttt{char}(S, \mu.)$.

Rethinking Manipulation-proof. Yan and Zhang [2022] first propose manipulation-proof auditing that primarily revolves around fully reconstructing the model, and defines the manipulation-proof subclass using a version space. However, this approach may overlook numerous other models that, while having a significant probability mass in areas where they disagree with the black-box model, exhibit similar behavior to the black-box model w.r.t. the property. In contrast, we propose to capture all those functions by defining only the essential information required for auditing.

Definition 9 (Fourier strategic manipulation-proof)**.** *Let* h *be a model that admits a Fourier expansion as in* $h = \sum_{S \subseteq [n]} \hat{h}(S)\psi_S$. *We say that an auditor* $\mathcal{A}$ *achieves optimal manipulation-proof for estimating a (distributional) property* μ *when* $\mathcal{A}$ *is a PAC-agnostic auditor (Definition 5) and outputs an exponential-size subclass of functions that satisfies* $\forall h, h' \in \mathcal{M}, \mathbb{P}\left(|\mu(h) - \mu(h')| \geq \epsilon\right) \leq \delta$.

Theorem 4 (Manipulation-proof of AFA)**.** AFA *achieves optimal manipulation-proof for estimating statistical parity with manipulation-proof subclass of size* 2^{n-2}.

Lower Bounds without Manipulation-proof. In the following, we propose a lower bound for yielding a PAC estimate of the statistical parity with no manipulation-proof constraint. Additionally, we assume the auditing algorithm can sequentially query the black-box model with informative queries. The proof is in Appendix C.4.

Theorem 5 (Lower bound without manipulation-proof)**.** *Let* $\epsilon \in (0, 1)$, $\delta \in (0, 1/2]$. *We aim to obtain* (ϵ, δ)*-PAC estimate of SP of model* $h \in \mathcal{H}$*, where the hypothesis class* $\mathcal{H}$ *has VC dimension* d. *For any auditing algorithm* $\mathcal{A}$*, there exists an adversarial distribution realizable by the model to audit such that with* $\tilde{\Omega}(\frac{\delta}{\epsilon^2})$ *samples,* $\mathcal{A}$ *outputs an estimate* $\hat{\mu}$ *of* $\mu_{\mathrm{GFair}}(h^*)$ *with* $\mathbb{P}[|\hat{\mu} - \mu_{\mathrm{GFair}}(h^*)| > \epsilon] > \delta$.

Table 2: Average estimation error for statistical parity across different ML models. '—' denotes when a method cannot scale to the model. The best method is in **bold**.

Dataset	COMPAS			Student			Drug		
Model	LR	MLP	RF	LR	MLP	RF	LR	MLP	RF
μCAL	0.312	—	—	—	—	—	—	—	—
`Uniform`	0.077	0.225	0.077	0.132	0.225	0.077	0.254	0.116	0.127
AFA	**0.006**	**0.147**	**0.006**	**0.030**	**0.147**	**0.006**	**0.220**	**0.040**	**0.120**

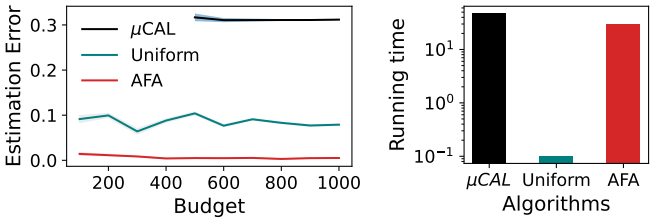


Figure 3: Error (left) and running time (right) of different auditors in estimating statistical parity of COMPAS in LR.

Table 3: Estimation error for robustness and individual fairness by `Uniform` and AFA. **Bold** case means lower error.

	Robustness		Individual Fairness	
ρ	`Uniform`	AFA	`Uniform`	AFA
0.25	0.033	**0.016**	0.036	**0.029**
0.30	0.333	**0.078**	0.309	**0.047**
0.35	0.299	**0.139**	0.248	**0.092**

Our results extend the existing sample complexity results with model reconstruction [Yan and Zhang, 2022], and also provide a reference of optimality for upper bounds. We highlight the gap from the upper bound established in Theorem 3, attributed to the lack of the manipulation proof.

5 Empirical Performance Analysis

In this section, we evaluate the performance of AFA in estimating multiple models' group fairness, robustness, and individual fairness. Below, we provide a detailed discussion of the experimental setup, objectives, and results.

Experimental Setup. We conduct experiments on COMPAS [Angwin et al., 2016], student performance (Student) [Cortez and Silva, 2008], and drug consumption (Drug) [Fehrman et al., 2019] datasets. The datasets contain a mix of binary, categorical, and continuous features for binary and multi-class classification.

We evaluate AFA on three ML models: Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF). The ground truth of group fairness, individual fairness, and robustness is computed using the entire dataset as in [Yan and Zhang, 2022].

For group fairness, we compare AFA with uniform sampling method, namely `Uniform`, and the active fairness auditing algorithms [Yan and Zhang, 2022, Algorithm 3], i.e. `CAL` and its variants μ`CAL` and randomized μ`CAL`, which requires more information about the model class than black-box access. We report the best variant of `CAL` with the lowest error. For robustness and individual fairness, we compare AFA with `Uniform`. Each experiment is run 10 times and we report the averages. We refer to Appendix E.1 for details.

Our empirical studies have the following **objectives**:
1. How accurate AFA is with respect to the baselines to audit robustness, individual fairness, and group fairness for different models and datasets?

2. How sample efficient and computationally efficient AFA is with baselines in auditing distributional properties?

Accurate, Sample Efficient, and Fast Estimation of Group Fairness.

In Table 2, we demonstrate the estimation error of group fairness by different methods across datasets and models. AFA yields the lowest estimation error, hence a better method, than all baselines in all nine configurations of models and datasets. Among baselines, `CAL` cannot estimate group fairness beyond COMPAS on LR, due to the requirement of a finite version space, which is provided only for COMPAS on LR. `Uniform`, albeit simple to implement, invariably demonstrates erroneous estimate. *Thus,* AFA *is the most accurate auditor for group fairness w.r.t. baselines.*

Figure 3 (left) demonstrates the sample efficiency of different methods for statistical parity. AFA requires the lowest number of samples to reach almost zero estimation error. Thus, AFA *is sample efficient than other methods*. Figure 3 (right) demonstrates the corresponding runtimes, where AFA is the second fastest method after `Uniform` and faster than `CAL`. *Therefore,* AFA *yields a well balance between accuracy, sample efficiency, and running time among baselines.*

Accurate Estimation of Robustness and Individual Fairness. Table 3 demonstrates the estimation error for robustness and individual fairness achieved by AFA and `Uniform` with different ρ's and 1000 samples from COMPAS dataset and LR model. AFA yields lower estimation error than `Uniform` across different models, and for higher values of ρ, the improvement due to AFA increases. Intuitively, `Uniform` samples IID from the space of input features, perturbs samples uniformly randomly, then queries the black-box model to obtain labels of perturbed samples to estimate properties. In contrast, AFA queries samples recursively to cover the feature space and estimates large Fourier coefficients without perturbing the input features. This also reflects the theoretical sample complexity results for `Uniform` and AFA, i.e. $O(1/\epsilon^2)$ and $O(1/\epsilon)$, respectively. *Thus,* AFA *is more accurate than* `Uniform` *to estimate robustness and individual fairness.*

6 Conclusion and Future Work

We propose AFA, a Fourier-based model-agnostic and black-box approach for universally auditing an ML model's distributional properties. We focus on three properties: robustness, individual fairness, and group fairness. We show that the significant Fourier coefficients of the black-box model yield a PAC approximation of all properties, establishing AFA as a universal auditor of ML. Empirically, AFA is more accurate, and sample efficient, while being competitive in running time than existing methods across datasets. In the future, we aim to extend AFA to estimate distributional properties other than the three studied in this paper.

7 Acknowledgements

This work was supported by the Regalia Project partnered by Inria and the French Ministry of Finance. D. Basu acknowledges the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01) and the PEPR project FOUNDRY (ANR23-PEIA-0003).

References

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. In *arxiv*, 2016. arxiv:1609.07236.

Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA): 1–30, 2017.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

David Cohn, Les Atlas, and Ladner Richard. Improving generalization with active learning. In *Machine Learning (20)*, pages 201–221, 1994.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 12)*, volume 86, pages 214—226, 2012.

Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multi-group fairness and back. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3566–3614. PMLR, 2023.

Elaine Fehrman, Vincent Egan, Alexander N Gorban, Jeremy Levesley, Evgeny M Mirkes, and Awaz K Muhammad. *Personality traits and drug consumption*. Springer, 2019.

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. Justicia: A stochastic sat approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7554–7563, 2021.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9539–9548, 2022.

Oded Goldreich and Leonid A. Levin. A hard-core predicate for all one-way functions. *In D. S. Johnson, editor, Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17*, 1989.

Shafi Goldwasser, Guy N. Rothblum, Jonathan Shafer, and Amir Yehudayoff. Interactive proofs for verifying machine learning. *Innovations in Theoretical Computer Science Conference (ITCS)*, 2021.

Mohsen Heidari, Jithin Sreedharan, Gil I Shamir, and Wojciech Szpankowski. Finding relevant information via a discrete fourier expansion. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139:4181-4191*, 2021.

Tal Herman and Guy N. Rothblum. Verifying the unseen: Interactive proofs for label-invariant distribution properties. *STOC: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 2022.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pages 1345–1362, 2020.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of the 35th International Conference on Machine Learning, PMLR*, 2018.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE security and privacy workshops (SPW)*, pages 69–75. IEEE, 2020.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2016.

Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. *SIAM J. Comput., 22(6):1331–1348*, 1993.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

Timothée Mathieu, Debabrota Basu, and Odalric-Ambrym Maillard. Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithms. *Transactions on Machine Learning Research*, 2022.

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.

Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021.

Saachi Mutreja and Jonathan Shafer. Pac verification of statistical algorithms. *36th Annual Conference on Learning Theory (COLT)*, 2023.

Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, Massachusetts, 2014.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

Khansa Rasheed, Adnan Qayyum, Mohammed Ghaly, Ala Al-Fuqaha, Adeel Razi, and Junaid Qadir. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 149:106043, 2022.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

Yifei Wang, Tavor Z Baharav, Yanjun Han, Jiantao Jiao, and David Tse. Beyond the best: Estimating distribution functionals in infinite-armed bandits. *arXiv preprint arXiv:2211.01743*, 2022.

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

Huan Xu and Mannor Shie. Robustness and generalization. In *Maéch Learn*, volume 86, pages 391—423, 2011.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pages 24929–24962. PMLR, 2022.

Appendix

Table of Contents

A The cost of auditing with reconstruction: Proof of Proposition 2

Proposition 2. *If $\hat{h}$ is the reconstructed model from h, then*

$$|\mu_{\mathrm{GFair}}(\hat{h}) - \mu_{\mathrm{GFair}}(h)| \leq \min\left\{1, \frac{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})]}{\min(\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = 1], \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = -1])}\right\}. \tag{4}$$

Proof. **Step 1.** We begin the proof by lower bounding the probability of yielding different predictions by h and $\hat{h}$.

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})] &= \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 0]\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[x_A = 0] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1]\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[x_A = 1] \\ &\geq p\Big(\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 0] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1]\Big)\end{aligned}$$

The first equality is a consequence of the law of total probability. The last inequality holds as we define $p \triangleq \min\{\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[x_A = 1], \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[x_A = 0]\}$.

Since $p \neq 0$, we get

$$\frac{1}{p}\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})] \geq \underbrace{\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 0]}_{\text{Term 1}} + \underbrace{\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1]}_{\text{Term 2}}. \tag{5}$$

Step 2. We observe that the Term 2 above can be rewritten as

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1|x_A = 1] &= \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1, h(\mathbf{x}) = -1|x_A = 1] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1, h(\mathbf{x}) = 1|x_A = 1] \\ &\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1|x_A = 1]\end{aligned}$$

The last inequality is true due to the fact that $\hat{h}(\mathbf{x}) = 1, h(\mathbf{x}) = -1$ is a sub-event of the event $h(x) \neq \hat{h}(x)$.

Now, by symmetry of h and $\hat{h}$, we get

$$\left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1|x_A = 1] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1|x_A = 1]\right| \leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1] \tag{6}$$

Similarly, working further with the Term 1 yields

$$\left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1|x_A = 0] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1|x_A = 0]\right| \leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 0] \tag{7}$$

Step 3. Finally, using triangle inequality yields

$$\begin{aligned}|\mu(\hat{h}) - \mu(h)| &\leq \left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1|x_A = 0] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1|x_A = 0]\right| + \left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) = 1|x_A = 1] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1|x_A = 1]\right| \\ &\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 0] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})|x_A = 1] \\ &\leq \frac{1}{p}\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})]\end{aligned}$$

The second step comes from inequalities (6) and (7), while the last one is due to inequality (5). □

B Computing Model's Properties with Fourier Coefficients: Proofs of Section 3.2

B.1 Robustness and Individual Fairness

Proposition 3. *Let $\rho \in [-1, 1]$.*

The robustness of a binary classifier $h : \{-1, 1\}^n \to \{-1, 1\}$ under the Γ_ρ flipping perturbation is equivalent to the ρ-flipping influence function, and thus, can be expressed as

$$\mu_{\text{Rob}}(h) = \text{Inf}_\rho(h) = \sum_{S \subseteq [n]} \rho^{|S|} \hat{h}(S)^2.$$

Proof.

Step 0: Robustness in terms of a composition of expectations over the perturbation Γ_ρ and $\mathcal{D}$. By the definition of robustness, we have:

$$\begin{aligned}\text{Inf}_\rho(h) &= \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim\Gamma_\rho(x)}} [h(\mathbf{x}) \neq h(\mathbf{y})] \\ &= \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim\Gamma_\rho(x)}} [h(\mathbf{x})h(\mathbf{y})]\end{aligned}$$

Where the second equation comes from the fact that h takes values in $\{-1, 1\}$.

Step 1: Robustness via operator approach. We commence the proof by defining the robust operator $\mathcal{T}_\rho : \{-1, 1\}^n \to \mathbb{R}$ as

$$\mathcal{T}_\rho h(\mathbf{x}) \triangleq \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})} [h(\mathbf{y})] .$$

Given the expression of the influence function in step 0, we have:

$$\begin{aligned}\text{Inf}_\rho(h) &= \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} [h(\mathbf{x})h(\mathbf{y})] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} [h(\mathbf{x}) \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})} h(\mathbf{y})] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} [h(\mathbf{x})\mathcal{T}_\rho h(\mathbf{x})] \\ &\triangleq \langle\, h, \mathcal{T}_\rho h \rangle_{\mathcal{D}}\end{aligned}$$

The second equation comes from the linearity of the expectation, and the last step comes from the definition of the inner product that depends on the distribution $\mathcal{D}$.

Now, we expand both the model h and the operator $\mathcal{T}_\rho h$ in the Gram-Schmidt basis given by Heidari et al. [2021]:

$$\begin{aligned}\text{Inf}_\rho(h) &= \langle \sum_{i_1=1}^{2^n} \hat{h}(S_{i_1})\psi_{S_{i_1}}(\cdot), \sum_{i_2=1}^{2^n} \hat{h}(S_{i_2}) \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\cdot)} [\psi_{S_{i_2}}(\mathbf{y})] \rangle_{\mathcal{D}} \\ &= \sum_{i_1=1}^{2^n} \sum_{i_2=1}^{2^n} \hat{h}(S_{i_1})\hat{h}(S_{i_2}) \langle\, \psi_{S_{i_1}}, f^\rho_{S_{i_2}} \rangle_{\mathcal{D}}\end{aligned}$$

Where, for all $\mathbf{x} \in \{-1, 1\}^n$ and for all $i \in \{1, \cdots, 2^n\}$, we used the following notation: $f^\rho_{S_i}(x) \triangleq \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(x)} [\psi_{S_i}(\mathbf{y})]$.

Step 2: Reduction of the robust operator to basis elements operators.

Let $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
f^{\rho}_{S_i}(\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})}[\psi_{S_i}(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})}[\chi_{S_i}(\mathbf{y})] - \sum_{j=1}^{i-1} \alpha_{i,j} \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})}[\psi_{S_j}(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(\mathbf{x})}[\prod_{k\in S_i} y_k] - \sum_{j=1}^{i-1} \alpha_{i,j} f^{\rho}_{S_j}(\mathbf{x}) \\
&= \prod_{k\in S_i} \mathop{\mathbb{E}}_{\mathbf{y}\sim N_\rho(x)}[y_k] - \sum_{j=1}^{i-1} \alpha_{i,j} f^{\rho}_{S_j}(\mathbf{x}) \\
&= \prod_{k\in S_i} \rho x_k - \sum_{j=1}^{i-1} \alpha_{i,j} f^{\rho}_{S_j}(\mathbf{x})
\end{aligned}$$

The second inequality comes from replacing each basis element from the Gram-Schmidt orthogonalization process with its expression in terms of parity functions. In the fourth equation, the expectation over each component is computed by the perturbation process Γ_ρ, that is for each k in S_i, x_k is flipped with probability $\frac{1-\rho}{2}$.

$$\begin{aligned}
f^{\rho}_{S_i}(\mathbf{x}) &= \rho^{|S_i|}\chi_{S_i}(\mathbf{x}) - \sum_{j=1}^{i-1} \alpha_{i,j} f^{\rho}_{S_j}(\mathbf{x}) \\
&= \rho^{|S_i|}\psi_{S_i}(\mathbf{x}) + \rho^{|S_i|} \sum_{j=1}^{i-1} \alpha_{i,j}\psi_{S_j}(\mathbf{x}) - \sum_{j=1}^{i-1} \alpha_{i,j} f^{\rho}_{S_j}(\mathbf{x}) \\
&= \rho^{|S_i|}\psi_{S_i}(\mathbf{x}) + \sum_{j=1}^{i-1} \alpha_{i,j}(\rho^{|S_i|}\psi_{S_j}(\mathbf{x}) - f^{\rho}_{S_j}(\mathbf{x}))\,.
\end{aligned}$$

Now, we compute the inner product left in step 1 to conclude the proof:

$$\begin{aligned}
\langle\, \psi_{S_i}, f^{\rho}_{S_k}\rangle_{\mathcal{D}} &= \langle\, \psi_{S_i}, \rho^{|S_k|}\psi_{S_k} + \sum_{j=1}^{k-1} \alpha_{k,j}(\rho^{|S_k|}\psi_{S_j} - f^{\rho}_{S_j})\rangle_{\mathcal{D}} \\
&= \rho^{|S_k|}\delta_{i,k} + \sum_{j=1}^{k-1} \delta_{j,k}\alpha_{k,j}(\rho^{|S_k|} - \langle\, \psi_{S_i}, f^{\rho}_{S_j}\rangle_{\mathcal{D}}) \\
\langle\, \psi_{S_i}, f^{\rho}_{S_k}\rangle_{\mathcal{D}} &= \rho^{|S_k|}\delta_{i,k}
\end{aligned}$$

Where the last step comes from the fact that $j < k$.

Step 4: Conclusion.

$$\begin{aligned}
\mathrm{Inf}_\rho(h) &= \sum_{i_1=1}^{2^n}\sum_{i_2=1}^{2^n} \hat{h}(S_{i_1})\hat{h}(S_{i_2})\langle\, \psi_{S_{i_1}}, f^{\rho}_{S_{i_2}}\rangle_{\mathcal{D}} \\
&= \sum_{i_1=1}^{2^n}\sum_{i_2=1}^{2^n} \hat{h}(S_{i_1})\hat{h}(S_{i_2})\rho^{|S_{i_1}|}\delta_{i_1,i_2}
\end{aligned}$$

$$= \sum_{i=1}^{2^n} \rho^{|S_i|} \hat{h}(S_i)^2$$
$$= \sum_{S \subseteq [n]} \rho^{|S|} \hat{h}(S)^2$$

We deduce the Fourier pattern in robustness property:

$$\mu_{\text{Rob}}(h) = \sum_{S \subseteq [n]} \rho^{|S|} \hat{h}(S)^2$$

□

The proof for individual fairness proceeds similarly by considering the operator $\mathcal{T}_\rho : \{-1, 1\}^n \to \mathbb{R}$, defined as:

$$\mathcal{T}_{\rho,l} h(\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim N_{\rho,l}(\mathbf{x})}[h(\mathbf{y})]$$

B.2 Group Fairness: Statistical Parity

We first establish the relationship between group fairness and Fourier coefficients.

Lemma 1. *If $\text{Inf}_A(h)$ denotes the membership influence function for the sensitive attribute A of the model h, we have the following result that relates the influence function to the model's h Fourier coefficients:*

$$\text{Inf}_A(h) = \sum_{\substack{S \subseteq [n] \\ S \ni A}} \hat{h}(S)^2$$

Proof. The membership influence function for the sensitive attribute A is given by:

$$\text{Inf}_A(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x} \sim \mathcal{D}^+ \\ \mathbf{y} \sim \mathcal{D}^-}} [h(\mathbf{x}) \neq h(\mathbf{y})]$$

This function is closely related to the Laplacian of the target model in the direction of the sensitive attribute A, defined as:

$$L_A h(\mathbf{x}, \mathbf{y}) := \frac{h(\mathbf{x}) - h(\mathbf{y})}{2}, \forall (\mathbf{x}, \mathbf{y}) \in (\mathcal{X}^+, \mathcal{X}^-)$$

Since h takes values in $\{-1, 1\}$, one can see that $|L_A h(\mathbf{x}, \mathbf{y})|^2 = \mathbb{1}_{\{h(\mathbf{x}) \neq h(\mathbf{y})\}}$.

By taking the expectation over the left and right part:

$$\|L_A h\|^2_{\mathcal{D}^+, \mathcal{D}^-} = \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim \mathcal{D}^+ \\ \mathbf{y} \sim \mathcal{D}^-}} [L_A h(\mathbf{x}, \mathbf{y})^2] = \text{Inf}_A(h)$$

$$\begin{aligned}
\forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^+ \times \mathcal{X}^- : L_A h(\mathbf{x}, \mathbf{y}) &= \frac{1}{2} \sum_{S \subseteq [n]} \hat{h}(S) \psi_S(\mathbf{x}) - \frac{1}{2} \sum_{S \subseteq [n]} \hat{h}(S) \psi_S(\mathbf{y}) \\
&= \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \ni A}} \hat{h}(S) \psi_S(\mathbf{x}) + \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \not\ni A}} \hat{h}(S) \psi_S(\mathbf{x}) \\
&\quad - \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \ni A}} \hat{h}(S) \psi_S(\mathbf{y}) - \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \not\ni A}} \hat{h}(S) \psi_S(\mathbf{y}) \\
\forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^+ \times \mathcal{X}^- : L_A h(\mathbf{x}, , \mathbf{y}) &= \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \ni A}} \hat{h}(S) \psi_S(\mathbf{x}) - \frac{1}{2} \sum_{\substack{S \subseteq [n] \\ S \ni A}} \hat{h}(S) \psi_S(\mathbf{y})
\end{aligned}$$

By Parseval identity, $\|L_A h\|^2_{\mathcal{D}^+,\mathcal{D}^-} = \sum_{\substack{S\subseteq[n]\\ S\ni A}} \hat{h}(S)^2$.

Hence,

$$\mathrm{Inf}_A(h) = \|L_A h\|^2_{\mathcal{D}^+,\mathcal{D}^-} = \sum_{\substack{S\subseteq[n]\\ S\ni A}} \hat{h}(S)^2$$

□

Proposition 5. *Statistical parity of h w.r.t a sensitive attribute A and distribution $\mathcal{D}$ is the root of the second order polynomial*

$$P_{\hat{h}}(X) \triangleq \alpha(1-\alpha)X^2 - \hat{h}(\emptyset)(1-2\alpha)X - \sum_{S\subseteq[n], S\ni A} \hat{h}(S)^2 - \frac{(1-\hat{h}^2(\emptyset))}{2}\,, \tag{8}$$

where $\alpha = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[x_A = 1]$ and $\hat{h}(\emptyset)$ is the coefficient of the empty set.

Proof. We use the following notation in the proof:

$$\begin{aligned}
p &= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1] \\
\alpha &= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathcal{X}^+] \quad \text{(probability of belonging to the first sensitive group)} \\
\mu^+_{\mathrm{GFair}}(h) &= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = 1] \\
\mu^-_{\mathrm{GFair}}(h) &= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = -1]
\end{aligned}$$

We have,

$$\mu_{\mathrm{GFair}}(h) = \mu^+_{\mathrm{GFair}}(h) - \mu^-_{\mathrm{GFair}}(h) \tag{9}$$

By the law of total probability, we also have:

$$p = \alpha\mu^+_{\mathrm{GFair}}(h) + (1-\alpha)\mu^-_{\mathrm{GFair}}(h) \tag{10}$$

We first express the membership influence function in terms of the statistical parity:

$$\begin{aligned}
\mathrm{Inf}_A(h) &= \mathbb{P}_{\mathbf{x},\mathbf{x}'\sim\mathcal{D}}[h(\mathbf{x}) \neq h(\mathbf{x}') | x_A = 1, x'_A = -1] \\
&= \mathbb{P}_{\mathbf{x},\mathbf{x}'\sim\mathcal{D}}[h(\mathbf{x}) = 1, h(\mathbf{x}') = 0 | x_A = 1, x'_A = -1] + \mathbb{P}_{\mathbf{x},\mathbf{x}'\sim\mathcal{D}}[h(\mathbf{x}) = -1, h(\mathbf{x}') = 1 | x_A = 1, x'_A = -1] \\
&= \mu^+_{\mathrm{GFair}}(h)(1 - \mu^-_{\mathrm{GFair}}(h)) + \mu^-_{\mathrm{GFair}}(h)(1 - \mu^+_{\mathrm{GFair}}(h)) \\
&= \mu^+_{\mathrm{GFair}}(h) + \mu^-_{\mathrm{GFair}}(h) - 2\mu^+_{\mathrm{GFair}}(h)\mu^-_{\mathrm{GFair}}(h)
\end{aligned}$$

Hence, we have:

$$\mu^+_{\mathrm{GFair}}(h) + \mu^-_{\mathrm{GFair}}(h) - 2\mu^+_{\mathrm{GFair}}(h)\mu^-_{\mathrm{GFair}}(h) - \mathrm{Inf}_A(h) = 0$$

From equation 9, and equation 10, we have:

$$\begin{cases}
\mu^+_{\mathrm{GFair}}(h) &= p + (1-\alpha)\mu_{\mathrm{GFair}}(h) \\
\mu^-_{\mathrm{GFair}}(h) &= p - \alpha\mu_{\mathrm{GFair}}(h)
\end{cases}$$

The expression becomes:

$$2\alpha(1-\alpha)\mu_{\mathrm{GFair}}(h)^2 + (1-2p)(1-2\alpha)\mu_{\mathrm{GFair}}(h) - \mathrm{Inf}_A(h) + 2p(1-p) = 0$$

The Fourier coefficient of the empty set is given by:

$$\begin{aligned}\hat{h}(\emptyset) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[2\mathbb{1}_{\{h(\mathbf{x})=1\}} - 1] \\ \hat{h}(\emptyset) &= 2\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1] - 1\end{aligned}$$

Since $p = \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1]$, we get the desired result.

□

Corollary 2. *If $\mathcal{D}$ is the uniform distribution, statistical parity is exactly the Fourier coefficient of the sensitive attribute, i.e.*

$$\mu_{\mathrm{GFair}}(h) = \hat{h}(\{A\})$$

Proof.

$$\begin{aligned}\mu_{\mathrm{GFair}}(h) &= |\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(x) = y|x \in A^+] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(x) = y|x \in A^-]| \\ &= |\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(x) = y|x \in A^+] - \frac{1}{2} - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(x) = y|x \in A^-] + \frac{1}{2}| \\ &= |\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[2\mathbb{1}_{\{h(\mathbf{x})=1\}} - 1|x \in A^+] - \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[2\mathbb{1}_{\{h(\mathbf{x})=1\}} - 1|x \in A^-]| \\ &= |\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})|x \in A^+] - \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})|x \in A^-]| \\ &= |\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\psi_A(x)|x \in A^+] - \frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\psi_A(x)|x \in A^-]|\end{aligned}$$

□

C Theoretical Analysis: Proofs of Section 4

C.1 Upper Bounds on Sample Complexity of AFA

Claim 1. *Let $\{A_i\}_{i\in\mathcal{I}}$ a finite set of events indexed by $\mathcal{I}$. Then,*

$$\mathbb{P}\left[\bigcap_{i\in\mathcal{I}} A_i\right] \geq \sum_{i\in\mathcal{I}} \mathbb{P}\left[A_i\right] - |\mathcal{I}| + 1$$

The proof is a consequence of the union bound.

Lemma 2 (Two-Sample Hoeffding's Inequality)**.** *If $X_1, ...X_{m_1}$, $X'_1, ...X'_{m_2}$ are iid random variables taking values in $[-1, 1]$ generating by the distribution $\mathcal{D}$, such that*

$$\mu = \mathbb{E}[X^2], \text{ and } \hat{\mu} = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} X_i X'_j,$$

then

$$\mathbb{P}[|\hat{\mu} - \mu| \leq 4\epsilon] \geq 1 - 2\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{8}\Big\}$$

The proof is obtained by employing one sample Hoeffding inequality to the random variable $Z_{i,j} = X_i X'_j$.

Theorem 2 (Upper bounds for Robustness and Individual Fairness)**.** *Given $\epsilon \in (0,1)$ and $\delta \in (0,1]$,* AFA *is a PAC-agnostic auditor for robustness and individual fairness with sample complexity*

$$\mathcal{O}\Big(\frac{\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\Big).$$

Here, $\texttt{char}(L,\mu.) \triangleq \sum_{S\in L}\texttt{char}(S,\mu.)$ *and* $\texttt{char}(\bar{L},\mu.) \triangleq \sum_{S\in \bar{L}}\texttt{char}(S,\mu.)$.

Proof.

Step 0. Let us define $\tau^2 \triangleq 4\epsilon$.

Let $x_1,\cdots,x_{m_1},x'_1,\cdots,x'_{m_2}$ are sampled i.i.d. from $\mathcal{D}$, where $m = m_1 + m_2$ denotes the total number of samples, and m_1 and m_2 to be the number of samples with $x_a = 1$ and $x_A = -1$, respectively.

Let L denote the list of subsets exhibiting Fourier coefficients larger than τ.

Step 1. By definitions of $\texttt{char}(S,\mu.)$, and the results of Proposition 3 and 4, we unifiedly express both the 'true' properties of h as

$$\begin{aligned}\mu(h) &= \sum_{S\subseteq[\![1,n]\!]}\texttt{char}(S,\mu)\hat{h}(S)^2\\ &= \sum_{S\subseteq[\![1,n]\!]}\texttt{char}(S,\mu)\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}}[h(\mathbf{x})h(\mathbf{y})\psi_S(\mathbf{x})\psi_S(\mathbf{y})]\,. \end{aligned}\tag{11}$$

Now, for any $S \in L$, we define an unbiased estimator of the squared Fourier coefficients as

$$\hat{h}_{\mathsf{AFA}}(S)^2 \triangleq \frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}h(x_i)h(x'_j)\psi_S(x_i)\psi_S(x_j)\,.\tag{12}$$

Hence, the estimators of these properties, i.e. robustness and individual fairness, takes the form

$$\hat{\mu}_{\mathsf{AFA}} \triangleq \frac{1}{m_1m_2}\sum_{S\in L}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\texttt{char}(S,\mu)h(x_i)h(x'_j)\psi_S(x_i)\psi_S(x_j)\,.\tag{13}$$

Step 2. Using Equation (11) and (13), we express the estimation error as

$$\begin{aligned}|\mu(h)-\hat{\mu}_{\mathsf{AFA}}| &= \left|\sum_{S\subseteq[\![1,n]\!]}\texttt{char}(S,\mu)\hat{h}(S)^2 - \sum_{S\in L}\texttt{char}(S,\mu)\hat{h}_{\mathsf{AFA}}(S)^2\right|\\ &= \left|\sum_{S\in L}\texttt{char}(S,\mu)\hat{h}(S)^2 + \sum_{S\notin L}\texttt{char}(S,\mu)\hat{h}(S)^2 - \sum_{S\in L}\texttt{char}(S,\mu)\hat{h}_{\mathsf{AFA}}(S)^2\right|\\ &\le \sum_{S\notin L}\texttt{char}(S,\mu)\hat{h}(S)^2 + \sum_{S\in L}\texttt{char}(S,\mu)\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\\ &\le \tau^2\sum_{S\notin L}\texttt{char}(S,\mu) + \sum_{S\in L}\texttt{char}(S,\mu)\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\,. \end{aligned}\tag{14}$$

The penultimate inequality is due to the fact that $|x+y| \le |x| + |y|$ for all $x, y \in \mathbb{R}$. The last inequality is by the definition of L, i.e. $\forall S \subseteq [\![1,n]\!] : S \notin L$ implies that $|\hat{h}(S)| \le \tau$. AFA gets access to this list of subsets L due to the Goldreich-Levin algorithm.

Step 3. Now, we leverage Equation (14), to derive an PAC estimation bound for robustness and individual fairness. Specifically,

$$\mathbb{P}\Big[|\mu(h)-\hat{\mu}_{\mathsf{AFA}}| \ge \epsilon\Big] \le \mathbb{P}\Big[\sum_{S\in L}\texttt{char}(S,\mu)|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2| \ge \epsilon - \tau^2\texttt{char}(\bar{L},\mu)\Big]\tag{15}$$

Here, we denote by $\texttt{char}(L,\mu)$ the sum $\sum_{S\in L}\texttt{char}(S,\mu)$ and $\texttt{char}(\bar{L},\mu)$ the sum $\sum_{S\notin L}\texttt{char}(S,\mu)$.

Step 4. Now, by consecutively applying Claim 1 and Lemma 2, we get an upper bound on the estimation error of the squared Fourier coefficients in L.

$$\begin{aligned}\mathbb{P}\Big[\bigcap_{S\in L}\Big\{\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\le 4\epsilon\Big\}\Big] &\ge \sum_{S\in L}\mathbb{P}\Big[\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\le 4\epsilon\Big]-|L|+1\\ &\ge |L|-2|L|\exp\Big\{-\frac{m_1m_2\epsilon^2}{8}\Big\}-|L|+1\\ &\ge 1-2|L|\exp\Big\{-\frac{m_1m_2\epsilon^2}{8}\Big\}\end{aligned}$$

This result naturally yields a bound on $\sum_{S\in L}\texttt{char}(S,\mu)\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|$.

$$\begin{aligned}\mathbb{P}\Big[\sum_{S\in L}\texttt{char}(S,\mu)\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\ge 4\texttt{char}(L,\mu)\epsilon\Big] &\le \mathbb{P}\Big[\bigcup_{S\in L}\Big\{\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\ge 4\epsilon\Big\}\Big]\\ &\le 2|L|\exp\Big\{-\frac{m_1m_2\epsilon^2}{8}\Big\}\end{aligned}$$

The last inequality is due to the union bound.

Step 5. Finally, using the fact that $4\epsilon=\tau^2$ and properly substituting to ensure $4\texttt{char}(L,\mu)\epsilon\ge\epsilon-\tau^2\texttt{char}(\bar{L},\mu)$, we get

$$\mathbb{P}\Big[\sum_{S\in L}\texttt{char}(S,\mu)\Big|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2\Big|\ge\epsilon-\tau^2\texttt{char}(\bar{L},\mu)\Big]\le 2|L|\exp\Big\{-\frac{m_1m_2\epsilon^2}{128\texttt{char}(L,\mu)^2(1-4\texttt{char}(\bar{L},\mu))^2}\Big\}$$

Hence, by Equation (15),

$$\mathbb{P}\Big[|\mu(h)-\hat{\mu}_{\mathsf{AFA}}|\ge\epsilon\Big]\le 2|L|\exp\Big\{-\frac{m_1m_2\epsilon^2}{128\texttt{char}(L,\mu)^2(1-4\texttt{char}(\bar{L},\mu))^2}\Big\}$$

By the definition of the sample complexity, the probability in the RHS has to be less than a given δ. Thus,

$$m_1m_2\ge\frac{128\texttt{char}(L,\mu)^2(1-4\texttt{char}(\bar{L},\mu))^2}{\epsilon^2}\log\frac{2|L|}{\delta}\,.$$

Since $L\ge 1$ and $m=m_1+m_2\ge 2\sqrt{m_1m_2}$, we conclude

$$m\ge\frac{8\sqrt{2}\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}$$

□

Theorem 3 (Upper bounds for Group Fairness)**.** *Given $\epsilon \in (0,1)$ and $\delta \in (0,1]$,* AFA *yields an (ϵ,δ)-PAC estimate of $\mu_{\mathrm{GFair}}(h)$ if it has access to predictions of*

$$\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{4}{\delta}\right)$$

input samples.

Proof. **Step 1.** First, we aim to express the group fairness as a root of the second-order polynomial in Proposition 5, and thus, to check when this approach is valid.

We observe that the discriminant of this second order polynomial is

$$\begin{aligned}
\Delta &= (2p+1)^2(2\alpha-1)^2 + 8\alpha(1-\alpha)\mathrm{Inf}_A - 1\\
&= 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8\alpha(1-\alpha)\mathrm{Inf}_A\\
&= 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8\alpha(1-\alpha)\sum_{S\subseteq[\![1,n]\!]}\hat{h}^2(S)\\
&= 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8\alpha(1-\alpha)\sum_{S\in L}\hat{h}^2(S) + 8\alpha(1-\alpha)\sum_{S\notin L}\hat{h}^2(S)\\
&\geq 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8\alpha(1-\alpha)\sum_{S\in L}\hat{h}^2(S)\\
&\geq 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8|L|\tau^2\alpha(1-\alpha)\\
&\geq 4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 32\epsilon\alpha(1-\alpha)\\
&= 4(1-8\epsilon)(\alpha-\frac{1}{2})^2 + 4(p-\frac{1}{2})^2 - (1-8\epsilon)
\end{aligned}$$

For $\epsilon > \frac{1}{8}$, Δ is positive. Thus, $\mu_{\mathrm{GFair}}(h)$, i.e. the zero of a second-order polynomial, can be expressed as

$$\mu_{\mathrm{GFair}}(h) = \frac{-(1-2\alpha)(1-2p) + \left(4\alpha^2 + 4p^2 - 4\alpha - 4p + 1 + 8\alpha(1-\alpha)\mathrm{Inf}_A\right)^{0.5}}{4\alpha(1-\alpha)}$$

Here, $p = \frac{1+\hat{h}(\emptyset)}{2}$ and $\mathrm{Inf}_A = \mathrm{Inf}_A(h) = \sum_{S\subseteq[\![1,n]\!],\ S\ni A}\hat{h}(S)^2$.

Step 2. We consider the following estimator yielded by AFA[1].

$$\hat{\mu}_{\mathrm{GFair}}(h) = \frac{-(1-2\alpha)(1-2\hat{p}) + \left(4\alpha^2 + 4\hat{p}^2 - 4\alpha - 4\hat{p} + 1 + 8\alpha(1-\alpha)\widehat{\mathrm{Inf}_A}\right)^{0.5}}{4\alpha(1-\alpha)},$$

where

$$\hat{p} = \frac{1+\hat{h}_{\mathsf{AFA}}(\emptyset)}{2}, \quad \text{and} \quad \widehat{\mathrm{Inf}_A} = \sum_{\substack{S\in L\\ S\ni A}}\hat{h}_{\mathsf{AFA}}(S)^2\,.$$

To simplify notations, we denote:

$$\Delta = 4\alpha^2 + 4p^2 - 4\alpha - 4p + 8\alpha(1-\alpha)\mathrm{Inf}_A + 1 \tag{16}$$

[1]Note that this estimator is independent of α or p, unlike the restrictive assumptions required in existing works [Yan and Zhang, 2022].

$$\hat{\Delta} = 4\hat{\alpha}^2 + 4\hat{p}^2 - 4\alpha - 4p + 8\alpha(1-\alpha)\widehat{\mathrm{Inf}}_A + 1 \tag{17}$$

Step 3. We have,

$$\mathbb{P}\left[|\widehat{\mu_{\mathrm{GFair}}} - \mu_{\mathrm{GFair}}(h)| \leq \epsilon\right] \geq \mathbb{P}\left[|\hat{p} - p| \leq \frac{2\alpha(1-\alpha)\epsilon}{|1-2\alpha|}\right] + \mathbb{P}\left[|\hat{\Delta} - \Delta| \leq 2\alpha(1-\alpha)\epsilon\right] - 1$$

On the other hand,

$$\mathbb{P}\left[|\hat{\Delta} - \Delta| \leq \epsilon\right] \geq \mathbb{P}\left[|\hat{p}^2 - p^2| \leq \frac{\epsilon}{12}\right] + \mathbb{P}\left[|\hat{p} - p| \leq \frac{\epsilon}{12}\right] + \mathbb{P}\left[|\widehat{\mathrm{Inf}}_A - \mathrm{Inf}_A| \leq \frac{\epsilon}{24\alpha(1-\alpha)}\right]$$

Similar to the previous proof and we apply using Two-sample Hoeffding on the first and third term above, while we use the classical Hoeffding for the second term. Together they yield a sample complexity upper bound of $\mathcal{O}\left(\max\left\{\frac{1}{\epsilon^2}\log\frac{4}{\delta}, \frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right\}\right)$, which is $\mathcal{O}(\frac{1}{\epsilon^2}\log\frac{4}{\delta})$ for $\epsilon \in (0,1)$ and $\delta \in (0,1]$. □

C.2 Time complexity

Here, we prove that AFA outputs (ϵ, δ)-PAC estimates of robustness, individual fairness and group fairness in time complexity poly $\left(\frac{1}{\epsilon}, n\right)$:

Given the graph shown in Figure 2, at any level of the tree, the set of active nodes, denoted by $\mathcal{N}_a$ correspond to the nodes whose weights are at least $\frac{\tau}{2}$, and hence the ones being processed by AFA.

We have,

$$|\mathcal{N}_a|\frac{\tau^2}{4} \leq \sum_{S \in \mathcal{N}_a} |\hat{h}(S)|^2$$

By Parseval identity,

$$\sum_{S \in \mathcal{N}_a} |\hat{h}(S)|^2 \leq 1$$

Hence,

$$|\mathcal{N}_a| \leq \frac{4}{\tau^2}$$

On the other hand, since we end up with subsets of a single element, the total number of splits [2] leading to the final nodes is at most n.

Hence the total number of estimates AFA performs is at most $\frac{8n}{\tau^2}$, which is poly $\left(\frac{1}{\epsilon}, n\right)$.

C.3 Manipulation-proof of AFA

Theorem 4 (Manipulation-proof of AFA)**.** AFA *achieves optimal manipulation-proof for estimating statistical parity with manipulation-proof subclass of size* 2^{n-2}.

Proof. We are interested in hypotheses h for which $\mu_{\mathrm{GFair}}(h) = \mu_{\mathrm{GFair}}(h^*)$.

[2] The total number of splits corresponds to the depth of the tree in Figure 2.

Let h^* denote the model under audit and let h be any model that admits Fourier decomposition, we have:

$$\begin{aligned} h &= \sum_{S\subseteq[n]} \hat{h}(S)\psi_S \\ &= \sum_{\substack{S\subseteq[n]\\ S\neq\emptyset}} \hat{h}(S)\psi_S + \hat{h}(\emptyset)\psi_\emptyset \\ &= \sum_{\substack{S\subseteq[n]\\ S\neq\emptyset, S\ni A}} \hat{h}(S)\psi_S + \sum_{\substack{S\subseteq[n]\\ S\neq\emptyset, S\not\ni A}} \hat{h}(S)\psi_S + \hat{h}(\emptyset)\psi_\emptyset \end{aligned}$$

On the other hand,

$$\forall S : S \ni A, \hat{h}(S) = h^*(S), \hat{h}(\emptyset) = h^*(\emptyset) \implies \mu_{\mathrm{GFair}}(h) = \mu_{\mathrm{GFair}}(h^*)$$

Where the last line comes from the dependence of statistical parity on the Fourier coefficients of the empty set and any subset that contains the protected feature (e.g, Formula 5).

Hence, the manipulation proof subclass is: $\left\{ h : \sum_{S\subseteq[n]} \hat{h}(S)\psi_S : \forall S \subseteq [n] : (S = \emptyset) \vee (S \ni A) \implies \hat{h}(S) = \hat{h^*}(S) \right\}$, which has a size of 2^{n-2}. □

C.4 Lower Bound on Sample Complexity without Manipulation-proof

Theorem 5 (Lower bound without manipulation-proof)**.** *Let $\epsilon \in (0,1)$, $\delta \in (0, 1/2]$. We aim to obtain (ϵ,δ)-PAC estimate of SP of model $h \in \mathcal{H}$, where the hypothesis class $\mathcal{H}$ has VC dimension d. For any auditing algorithm $\mathcal{A}$, there exists an adversarial distribution realizable by the model to audit such that with $\tilde{\Omega}(\frac{\delta}{\epsilon^2})$ samples, $\mathcal{A}$ outputs an estimate $\hat{\mu}$ of $\mu_{\mathrm{GFair}}(h^*)$ with $\mathbb{P}[|\hat{\mu} - \mu_{\mathrm{GFair}}(h^*)| > \epsilon] > \delta$.*

Proof. Let $\mathcal{H}$ be a hypothesis class of VC dimension $\mathrm{VC}(\mathcal{H})$, we start with case $\mathrm{VC}(\mathcal{H}) \in 2\mathbb{N}$.

Let $\mathcal{Z} = \{\zeta_1, \ldots, \zeta_d, \zeta_{d+1}, \cdots, \zeta_{2d}\} \subseteq \mathcal{X}$ a subspace shattered by $\mathcal{H}$, let N be our querying budget.

Step 1: Construction of adversarial distribution. Let $\mathcal{Z}^+ = \{\zeta_1, \ldots, \zeta_d\}$ and $\mathcal{Z}^- = \{\zeta_{d+1}, \ldots, \zeta_{2d}\}$.

We define the adversarial distribution as the distribution satisfying:

$$\mathcal{D} = \begin{cases} x|\mathcal{X}^+ & \sim \mathcal{U}\{\mathcal{Z}^+\} \\ x|\mathcal{X}^- & \sim \mathcal{U}\{\mathcal{Z}^-\} \end{cases}$$

For any $i \in [\![1, 2d]\!]$ and given the iid assumption, any $z \sim \mathcal{Z}^+$ will be denoted z^+ and similarly any $z \sim \mathcal{Z}^-$ will be denoted z^-.

Consider hypotheses H_0 and H_1 that chooses h^* randomly from $\{0,1\}^{\mathcal{Z}}$:

- H_0: picks h^* such that for all $i \in [\![1, d]\!]$ independently:

$$h^*(z_i) := \begin{cases} 1 & \text{with probability } \frac{1}{2} - \epsilon \\ 0 & \text{with probability } \frac{1}{2} + \epsilon \end{cases} \tag{18}$$

  and for all $i \in [\![d+1, 2d]\!]$ (independently):

$$h^*(z_i) := \begin{cases} 1 & \text{with probability } \frac{1}{2} + \epsilon \\ 0 & \text{with probability } \frac{1}{2} - \epsilon \end{cases} \tag{19}$$

- H_1: picks h^* such that for all $i \in [\![1, d]\!]$ independently:

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} + \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} - \epsilon \end{cases} \tag{20}$$

and for all $i \in [\![d+1, 2d]\!]$ (independently):

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} - \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} + \epsilon \end{cases} \tag{21}$$

If h^* is chosen under hypothesis H_i, the probability that involves h^* will be denoted $\mathbb{P}_i$.

The case where $\mathrm{VC}(\mathcal{H}) \in 2\mathbb{N}+1$ reduces to $\mathrm{VC}(\mathcal{H}) \in 2\mathbb{N}$ by giving a delta mass distribution to ζ_{2d+1} on the subspace shattered by $\mathcal{H}$.

Step 2: Bounding demographic parity by bounding p and Inf_A

In order to get a lower bound for estimating statistical parity, we express it in terms of the probability of positives and the randomized influence function.

$$\mathbb{P}\left[\hat{\mu} - \mu(h^*) > \epsilon\right] \geq \underbrace{\mathbb{P}\left[\hat{p} - p(h^*) > c_\alpha \epsilon\right]}_{\text{Term I}} + \underbrace{\mathbb{P}\left[\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*) > c_{\alpha,1}\epsilon^2 + c_{\alpha,2}\right]}_{\text{Term II}}, \tag{22}$$

where $c_\alpha = \frac{4\alpha(1-\alpha)(12-\sqrt{6})}{11(1-2\alpha)}$, $c_{\alpha,1} = 1 + \frac{1}{2\alpha(1-\alpha)}$ and $c_{\alpha,2} = \sqrt{\frac{2}{3}}\frac{1}{2\alpha^2(1-\alpha)^2}$.

Step 2.a: Bounding the Term I. Turning an estimation problem into a testing problem. Under hypothesis H_0, we have:

$$\begin{aligned}
\mathbb{P}_0\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] &\geq \mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha - 11}{2}, p(h^*) \leq \frac{2\alpha - 1}{2} - \frac{\epsilon}{2}\right] \\
&\geq \mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha - 1}{2}\right] + \mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha - 1}{2} - \frac{\epsilon}{2}\right] - 1
\end{aligned}$$

Under hypothesis H_1, we have:

$$\begin{aligned}
\mathbb{P}_1\left[\hat{p} - p(h^*) \leq \frac{\epsilon}{2}\right] &\geq \mathbb{P}_1\left[\hat{p} \geq \frac{2\alpha - 1}{2}, p(h^*) \geq \frac{2\alpha - 1}{2} + \frac{\epsilon}{2}\right] \\
&\geq \mathbb{P}_1\left[\hat{p} < \frac{2\alpha - 1}{2}\right] + \mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha - 1}{2} + \frac{\epsilon}{2}\right] - 1
\end{aligned}$$

Since

$$\mathbb{P}\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] = \frac{1}{2}\mathbb{P}_0\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] + \frac{1}{2}\mathbb{P}_1\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right]$$

Using the fact that $\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$, we have:

$$\mathbb{P}\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] \geq \frac{1}{2}\Bigg(\mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha - 1}{2}\right] + \mathbb{P}_1\left[\hat{p} < \frac{2\alpha - 1}{2}\right] \tag{23}$$

$$+ \mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha - 1}{2} - \frac{\epsilon}{2}\right] + \mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha - 1}{2} + \frac{\epsilon}{2}\right] - 2\Bigg) \tag{24}$$

By Le Cam's lemma:

$$\mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha - 1}{2}\right] + \mathbb{P}_1\left[\hat{p} < \frac{2\alpha - 1}{2}\right] \geq 1 - \mathrm{TV}\left(\mathbb{P}_0 \parallel \mathbb{P}_1\right) \tag{25}$$

Concentration of $p(h^*)$. To lower bound the remaining term in (24), we prove Lemma 3.

This proves result 31.

Similar to the proof of the first result, by Hoeffding inequality,

$$\mathbb{P}_1\left[p^+(h^*) > \frac{\alpha}{2} - \frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2\alpha^2}\Big)$$
$$\mathbb{P}_1\left[p^-(h^*) > \frac{1-\alpha}{2} - \frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big)$$

The proof of result 32 concludes by proceeding with the remaining steps in the same manner as the previous proof.

$$\mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{26}$$

$$\mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha-1}{2} + \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{27}$$

By symmetry of the statistical test we have the result in 32.

Step 2.b: Bounding Term II. Similar to **step 2.a**, we have:

$$\mathbb{P}\left[|\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*)| \geq \frac{\epsilon}{2}\right] = \frac{1}{2}\mathbb{P}_0\left[|\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*)| \geq \frac{\epsilon}{2}\right] + \frac{1}{2}\mathbb{P}_1\left[|\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*)| \geq \frac{\epsilon}{2}\right]$$

We deduce

$$\mathbb{P}\left[\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*) \geq \frac{\epsilon}{2}\right] \geq \frac{1}{2}\Bigg(\mathbb{P}_0\left[\widehat{\mathrm{Inf}_A} \geq \frac{1}{2}\right] + \mathbb{P}_1\left[\widehat{\mathrm{Inf}_A} < \frac{1}{2}\right] + \tag{28}$$

$$\mathbb{P}_0\left[\mathrm{Inf}_A(h^*) \leq \frac{1}{2} - \frac{\epsilon}{2}\right] + \mathbb{P}_1\left[\mathrm{Inf}_A(h^*) \geq \frac{1}{2} + \frac{\epsilon}{2}\right] - 2\Bigg) \tag{29}$$

By Le Cam's lemma, we have:

$$\mathbb{P}_0\Bigg(\mathrm{Inf}_A(h^*) > \frac{1}{2}\Bigg) + \mathbb{P}_1\Bigg(\mathrm{Inf}_A(h^*) \leq \frac{1}{2}\Bigg) \geq 1 - \mathrm{TV}\left(\mathbb{P}_0 \parallel \mathbb{P}_1\right)$$

Concentration of $\mathrm{Inf}_A(h^*)$. To lower bound the remaining term in (29), we prove Lemma 4.

Under hypothesis H_0, we have:

$$\begin{aligned}\mathbb{P}_0\left[\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*) \geq \frac{\epsilon^2}{2}\right] &\geq \mathbb{P}_0\left[\widehat{\mathrm{Inf}_A} \geq \frac{1}{2}, \mathrm{Inf}_A(h^*) \leq \frac{1}{2} - \frac{\epsilon^2}{2}\right] \\ &\geq \mathbb{P}_0\left[\widehat{\mathrm{Inf}_A} \geq \frac{1}{2}\right] + \mathbb{P}_0\left[\mathrm{Inf}_A(h^*) \leq \frac{1}{2} - \frac{\epsilon^2}{2}\right] - 1\end{aligned}$$

Under hypothesis H_1, we have:

$$\begin{aligned}\mathbb{P}_1\left[\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*) \geq \frac{\epsilon}{2}\right] &\geq \mathbb{P}_1\left[\widehat{\mathrm{Inf}_A} \geq \frac{1}{2}, \mathrm{Inf}_A(h^*) \geq \frac{1}{2} - \frac{\epsilon}{2}\right] \\ &\geq \mathbb{P}_1\left[\widehat{\mathrm{Inf}_A} < \frac{1}{2}\right] + \mathbb{P}_1\left[\mathrm{Inf}_A(h^*) \geq \frac{1}{2} - \frac{\epsilon}{2}\right] - 1\end{aligned}$$

Step 3: Upper bounding the statistical distances Let's show that H_0 and H_1 are hard to distinguish. In other words, let's show that $\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1) = \mathcal{O}(\epsilon^2)$

The quantity $\mathcal{D}_{KL}\Big(\mathbb{P}_0||\mathbb{P}_1\Big)$ depends on how the algorithm $\mathcal{A}$ interacts with the oracle $\mathcal{O}(h^*)$ and construct a brick of history denoted by $\mathcal{H}^{ist}$. We can observe that this quantity is exactly $\mathcal{D}_{KL}\Big(\mathbb{P}_0(y|(x,y) \in \mathcal{H}^{ist}, x)||\mathbb{P}_1(y|(x,y) \in \mathcal{H}^{ist}, x)\Big)$ averaged on the whole available querying set. More formally, we prove Lemma 5 that states

$$\mathcal{D}_{KL}\Big(\mathbb{P}_0||\mathbb{P}_1\Big) = \sum_{i=1}^{N} \mathbb{E}\left[\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\right)\right].$$

The next step is to upper bound this quantity: At iteration I, we distinguish between two separate cases:

- If $x_i \in \mathcal{H}_{i-1}^{ist}$, then $\mathcal{A}$ will always output the same value under both hypotheses H_0 and H_1, which was sent by oracle $\mathcal{O}(h^*)$. Hence,

$$D_{\mathrm{KL}}\left(\mathbb{P}_0(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\|\mathbb{P}_1(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\right) = 0$$

- If $x_i \notin \mathcal{H}_{i-1}^{ist}$, we have the following table that summarizes all possibilities under hypotheses H_0 and H_1, conditioning on $\mathcal{X}^+$:

$H \backslash y$	1	0
H_0	$\frac{1}{2} - \frac{\epsilon}{2}$	$\frac{1}{2} + \frac{\epsilon}{2}$
H_1	$\frac{1}{2} + \frac{\epsilon}{2}$	$\frac{1}{2} - \frac{\epsilon}{2}$

And under hypotheses H_0 and H_1, conditioning on $\mathcal{X}^-$:

$H \backslash y$	1	0
H_0	$\frac{1}{2} + \frac{\epsilon}{2}$	$\frac{1}{2} - \frac{\epsilon}{2}$
H_1	$\frac{1}{2} - \frac{\epsilon}{2}$	$\frac{1}{2} + \frac{\epsilon}{2}$

From the two tables, we deduce the overall result by expanding over each protected group (e.g, $\mathcal{X}^-, \mathcal{X}^+$)

$H \backslash y$	1	0
H_0	$\frac{1}{2} + \frac{(1-2\alpha)\epsilon}{2}$	$\frac{1}{2} - \frac{(1-2\alpha)\epsilon}{2}$
H_1	$\frac{1}{2} - \frac{(1-2\alpha)\epsilon}{2}$	$\frac{1}{2} + \frac{(1-2\alpha)\epsilon}{2}$

We end up with a binary entropy upper bound:

$$\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\right) = kl\left(\frac{1}{2} + \frac{(1-2\alpha)\epsilon}{2}, \frac{1}{2} - \frac{(1-2\alpha)\epsilon}{2}\right)$$

Fact 1. *For* $a, b \in (\frac{1}{4}, \frac{3}{4})$: $\mathcal{D}_{\mathrm{KL}}(a,b) \leq 3(b-a)^2$

Hence,

$$\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y) \in \mathcal{H}_{i-1}^{ist}, x_i)\right) \leq 3(1-2\alpha)^2\epsilon^2$$

$$\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1) \leq 3N(1-2\alpha)^2\epsilon^2 \tag{30}$$

By Pinsker's inequality;

$$\begin{cases} \mathbb{P}_0\Big(p(h^*) > \frac{1}{2}\Big) + \mathbb{P}_1\Big(p(h^*) \leq \frac{1}{2}\Big) \geq 1 - \sqrt{\frac{1}{2}\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1)} \\ \mathbb{P}_0\Big(\mathrm{Inf}_A(h^*) > \frac{1}{2}\Big) + \mathbb{P}_1\Big(\mathrm{Inf}_A(h^*) \leq \frac{1}{2}\Big) \geq 1 - \sqrt{\frac{1}{2}\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1)} \end{cases}$$

By using result from (30),

$$\begin{cases} \mathbb{P}_0\Big(p(h^*) > \frac{1}{2}\Big) + \mathbb{P}_1\Big(p(h^*) \leq \frac{1}{2}\Big) \geq 1 - \sqrt{\frac{3N(1-2\alpha)^2\epsilon^2}{2}} \\ \mathbb{P}_0\Big(\mathrm{Inf}_A(h^*) > \frac{1}{2}\Big) + \mathbb{P}_1\Big(\mathrm{Inf}_A(h^*) \leq \frac{1}{2}\Big) \geq 1 - \sqrt{\frac{3N(1-2\alpha)^2\epsilon^2}{2}} \end{cases}$$

Results in (31) and (32) further yield

$$\begin{aligned} \mathbb{P}\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] &\geq \frac{1}{2} - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) - \sqrt{\frac{3N}{2}}\frac{|1-2\alpha|\epsilon}{2} \\ &\geq \frac{1}{2} - 4\exp\Big(-\frac{d\epsilon^2}{8M_\alpha^2}\Big) - \sqrt{\frac{3N}{2}}\frac{|1-2\alpha|\epsilon}{2}, \end{aligned}$$

where $M_\alpha = \max(\alpha, 1-\alpha)$.

Further, (33) and (34) yield

$$\begin{aligned} \mathbb{P}\left[\widehat{\mathrm{Inf}_A} - \mathrm{Inf}_A(h^*) \geq \frac{\epsilon}{2}\right] &\geq \frac{5}{2} - 4\exp\Big(-\frac{d\epsilon}{2}\Big) - 4\exp\Big(-\frac{d\epsilon}{18}\Big) - \sqrt{\frac{3N(1-2\alpha)^2\epsilon^2}{8}} \\ &\geq \frac{5}{2} - 8\exp\Big(-\frac{d\epsilon}{18}\Big) - \sqrt{\frac{3N(1-2\alpha)^2\epsilon^2}{8}} \end{aligned}$$

Finally, solving the inequality

$$3 - 4\exp\frac{-d\epsilon^2}{18} - \sqrt{\frac{3N(1-2\alpha)^2\epsilon^2}{8}} \geq \delta$$

yields the sample complexity to be $N \leq \frac{8}{3(1-2\alpha)^2\epsilon^2}\left(\delta - 3 + 4\exp\left(-\frac{d\epsilon^2}{18}\right)\right)^2$. □

C.5 Additional Technical Lemmas

Lemma 3.

$$\mathbb{P}_0\Big[p(h^*) \le \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\Big] \ge 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{31}$$

$$\mathbb{P}_1\Big[p(h^*) \ge \frac{2\alpha-1}{2} + \frac{\epsilon}{2}\Big] \ge 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{32}$$

Proof.

$$\begin{aligned}
p(h^*) &= \mathbb{P}\Big[h^*(x) = 1\Big] \\
&= \alpha\,\mathbb{P}\Big[h^*(x) = 1\Big|\mathcal{X}^+\Big] + (1-\alpha)\,\mathbb{P}\Big[h^*(x) = 1\Big|\mathcal{X}^-\Big] \\
&= \frac{\alpha}{d}\sum_{i=1}^{d}\mathbb{1}_{\{h^*(z_i)=1\}} + \frac{1-\alpha}{d}\sum_{i=1}^{d}\mathbb{1}_{\{h^*(z_{d+i})=1\}} \\
p(h^*) &= p^+(h^*) + p^-(h^*)
\end{aligned}$$

Where $p^+(h^*) = \frac{\alpha}{d}\sum_{i=1}^{d}\mathbb{1}_{\{h^*(z_i)=1\}}$ and $p^-(h^*) = \frac{1-\alpha}{d}\sum_{i=1}^{d}\mathbb{1}_{\{h^*(z_{d+i})=1\}}$

Under H_0 (resp. H_1), $\frac{d}{\alpha}p^+(h^*)$ is the sum of d Bernoulli variables of mean $\frac{1}{2}-\epsilon$ (resp. $\frac{1}{2}+\epsilon$).
Under H_0 (resp. H_1), $\frac{d}{1-\alpha}p^+(h^*)$ is the sum of d Bernoulli variables of mean $\frac{1}{2}+\epsilon$ (resp. $\frac{1}{2}-\epsilon$).

$$\begin{aligned}
\mathbb{P}_0\Big[p^+(h^*) > \frac{\alpha}{2} - \frac{\epsilon}{4}\Big] &\le 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) \\
\mathbb{P}_0\Big[p^-(h^*) > \frac{\epsilon}{2} - \frac{1-\alpha}{2}\Big] &\le 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}_0\Big[p(h^*) \le \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\Big] &\ge \mathbb{P}_0\Big[p^+(h^*) \le \frac{\alpha}{2} - \frac{\epsilon}{4}, p^-(h^*) \le \frac{\epsilon}{2} - \frac{1-\alpha}{2}\Big] \\
&\ge \mathbb{P}_0\Big[p^+(h^*) \le \frac{\alpha}{2} - \frac{\epsilon}{4}\Big] + \mathbb{P}_0\Big[p^-(h^*) \le \frac{\epsilon}{2} - \frac{1-\alpha}{2}\Big] - 1 \\
&\ge 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big)
\end{aligned}$$

□

Lemma 4 (Concentration of Influence Function)**.**

$$\mathbb{P}_0\Big[\mathrm{Inf}_A(h^*) \le \frac{1+\epsilon}{2}\Big] \ge 3 - 4\exp\Big(-\frac{d\epsilon}{2}\Big) - 4\exp\Big(-\frac{d\epsilon}{18}\Big) \tag{33}$$

$$\mathbb{P}_1\Big[\mathrm{Inf}_A(h^*) > \frac{1-\epsilon}{2}\Big] \ge 3 - 4\exp\Big(-\frac{d\epsilon}{2}\Big) - 4\exp\Big(-\frac{d\epsilon}{18}\Big) \tag{34}$$

Proof.

$$\begin{aligned}
\mathrm{Inf}_A(h^*) &= \mathbb{P}\Big[h^*(x) \ne h^*(x')\Big|x \in \mathcal{X}^+, x' \in \mathcal{X}^-\Big] \\
&= \mathbb{P}\Big[h^*(x) = 1, h^*(x') = 0\Big|x \in \mathcal{X}^+, x' \in \mathcal{X}^-\Big] + \mathbb{P}\Big[h^*(x) = 0, h^*(x') = 1\Big|x \in \mathcal{X}^+, x' \in \mathcal{X}^-\Big]
\end{aligned}$$

$$= \mathbb{P}\left[h^*(x)=1 \middle| x \in \mathcal{X}^+\right]\mathbb{P}\left[h^*(x)=0 \middle| x \in \mathcal{X}^-\right] + \mathbb{P}\left[h^*(x)=0 \middle| x \in \mathcal{X}^+\right]\mathbb{P}\left[h^*(x)=1 \middle| x \in \mathcal{X}^-\right]$$
$$= \frac{1}{d^2}\sum_{1\leq i,j\leq d} \mathbb{1}_{\{h^*(z_i)=1\}}\mathbb{1}_{\{h^*(z_{d+j})=0\}} + \frac{1}{d^2}\sum_{1\leq i,j\leq d} \mathbb{1}_{\{h^*(z_i)=0\}}\mathbb{1}_{\{h^*(z_{d+j})=1\}}$$

$$\mathrm{Inf}_A(h^*) = \mathrm{Inf}^+_{A,1}(h^*)\mathrm{Inf}^-_{A,0}(h^*) + \mathrm{Inf}^+_{A,0}(h^*)\mathrm{Inf}^-_{A,1}(h^*)$$

Where, $\mathrm{Inf}^+_{A,1}(h^*) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}_{\{h^*(z_i)=1\}}$

$\mathrm{Inf}^-_{A,0}(h^*) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}_{\{h^*(z_{d+i})=0\}}$,

$\mathrm{Inf}^+_{A,0}(h^*) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}_{\{h^*(z_i)=0\}}$,

$\mathrm{Inf}^-_{A,1}(h^*) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}_{\{h^*(z_{d+i})=1\}}$.

- Under H_0 (resp. H_1), $\mathrm{Inf}^+_{A,1}(h^*)$) is the sum of d Bernoulli variables of mean $\frac{1}{2}-\epsilon$ (resp. $\frac{1}{2}+\epsilon$).
- Under H_0 (resp. H_1), $\mathrm{Inf}^-_{A,0}(h^*)$ is the sum of d Bernoulli variables of mean $\frac{1}{2}-\epsilon$ (resp. $\frac{1}{2}+\epsilon$).
- Under H_0 (resp. H_1), $\mathrm{Inf}^+_{A,0}(h^*)$ is the sum of d Bernoulli variables of mean $\frac{1}{2}+\epsilon$ (resp. $\frac{1}{2}-\epsilon$).
- Under H_0 (resp. H_1), $\mathrm{Inf}^-_{A,1}(h^*)$ is the sum of d Bernoulli variables of mean $\frac{1}{2}+\epsilon$ (resp. $\frac{1}{2}-\epsilon$).

Applying Hoeffding inequality under hypothesis H_0 gives:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,1}(h^*) > \frac{1}{2}-\frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{35}$$

$$\mathbb{P}_0\left[\mathrm{Inf}^-_{A,0}(h^*) > \frac{1}{2}-\frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{36}$$

From 35 and 36, we deduce:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,1}(h^*)\mathrm{Inf}^-_{A,0}(h^*) \leq \Big(\frac{1}{2}-\frac{\epsilon}{2}\Big)^2\right] \geq 2-4\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{37}$$

Similar, the upper bound of the second part is:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,0}(h^*)\mathrm{Inf}^-_{A,1}(h^*) \leq \Big(\frac{1}{2}+\frac{\epsilon}{2}\Big)^2\right] \geq 2-4\exp\Big(-\frac{d\epsilon^2}{18}\Big) \tag{38}$$

Combining results 37 and38 yields result 33. By the symmetry of the hypotheses H_0 and H_1, we obtain the second result. □

Lemma 5.

$$\mathcal{D}_{KL}\Big(\mathbb{P}_0||\mathbb{P}_1\Big) = \sum_{i=1}^N \mathbb{E}\left[\mathcal{D}_{KL}\Big(\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\Big\|\mathbb{P}_1(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\Big)\right]$$

Proof. By definition,

$$\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1) = \sum_{\mathcal{Q}\in\mathcal{H}_N^{ist}} \mathbb{P}_0(\mathcal{Q})\,\log\frac{\mathbb{P}_0(\mathcal{Q})}{\mathbb{P}_1(\mathcal{Q})}$$

$$= \sum_{\substack{\mathcal{Q}\in\mathcal{H}_N^{ist}\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_N,y_N)\}}} \mathbb{P}_0(\mathcal{Q})\,\log\frac{\prod_{i=1}^N \mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\mathbb{P}_{\mathcal{A}}(x_i|(x,y)\in\mathcal{H}_{i-1}^{ist})}{\prod_{i=1}^N \mathbb{P}_1((y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\mathbb{P}_{\mathcal{A}}(x_i|(x,y)\in\mathcal{H}_{i-1}^{ist})}$$

$$= \sum_{\substack{\mathcal{Q}\in\mathcal{H}_N^{ist}\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_N,y_N)\}}} \mathbb{P}_0(\mathcal{Q})\,\sum_{i=1}^N \log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}$$

$$= \sum_{i=1}^N \sum_{\substack{\mathcal{Q}\in\mathcal{H}_N^{ist}\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_i,y_i)\}}} \mathbb{P}_0(\mathcal{Q})\,\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}$$

$$= \sum_{i=1}^N \sum_{\{(x_1,y_1),\dots(x_i,y_i)\}} \mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\mathbb{P}_0((x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\,\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}$$

$$= \sum_{i=1}^N \sum_{\{(x_1,y_1),\dots(x_{i-1},y_{i-1}),x_i\}} \mathbb{P}_0((x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\sum_{y_i}\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)\,\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}_{i-1}^{ist},x_i)}$$

$$= \sum_{i=1}^N \sum_{\mathbf{H}_{i-1},x_i} \mathbb{P}_0((x,y)\in\mathbf{H}_{i-1},x_i)\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\right)$$

Hence,

$$\mathcal{D}_{KL}\Big(\mathbb{P}_0||\mathbb{P}_1\Big) = \sum_{i=1}^N \mathbb{E}\left[\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\right)\right]$$

□

D Extensions to Multi-class Classification

If $\{a_1,\cdots,a_n\}$ denotes the set of categories such that for all $i\neq j\in\{1,\cdots,n\}, \mathcal{X}_{i,j} = h^{-1}(\{a_i,a_j\})$, and A_h the set:

$$\mathcal{A}_h = \bigcup_{i\neq j}\Big\{h_{i,j}:\mathcal{X}_{i,j}\to\{a_i,a_j\}, h_{i,j}(\mathcal{X}_{i,j}) = h(\mathcal{X}_{i,j})\Big\}$$

Based on the result in Proposition 5 the Fourier pattern of multicalibration is as follows:

$$\mu_{\text{Rob}}(h) = \max_{g\in\mathcal{A}_h} P_{\hat{g}}^{-1}(0)$$

This adaptation is evaluated empirically to assess how well `AFA` performs in this setting.

E Experimental Details

All our computations are performed on an 11th Gen Intel® Core™ i7-1185G7 processor (3.00 GHz, 8 cores) with 32.0 GiB of RAM.

E.1 Uniformly Random Sampling (I.I.D.) estimators (`Uniform`)

Random estimators use i.i.d. sampling in order to estimate each distributional property. We note that group fairness estimation requires a different sampling strategy and interaction with the black-box oracle of h.

Robustness. The true robustness is defined as:

$$\mu_{\mathrm{Rob}}(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} [h(\mathbf{x}) \neq h(\mathbf{y})]$$

Random estimator samples i.i.d. points from $\mathcal{D}$, which we denote as S. Thus, the estimator can be written as

$$\widehat{\mu_{\mathrm{Rob}}(h)} = \frac{1}{|S|} \sum_{\substack{\mathbf{x}\in S\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Individual Fairness. Likewise, individual fairness estimation given by random estimator is:

$$\widehat{\mu_{\mathrm{IFair}}(h)} = \frac{1}{|S|} \sum_{\substack{\mathbf{x}\in S\\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Group Fairness. Let S^+ denote a set of samples from the first protected group and S^- a set of samples from the second protected group. Group Fairness (with demographic parity measure) is defined as:

$$\widehat{\mu_{\mathrm{GFair}}(h)} = \frac{1}{|S^+|} \sum_{\mathbf{x}\in S^+} \mathbb{1}_{h(\mathbf{x})=1} - \frac{1}{|S^-|} \sum_{\mathbf{x}\in S^-} \mathbb{1}_{h(\mathbf{x})=1}$$

E.2 Baseline Algorithms

We assess AFA on statistical parity by comparing its performance in sample complexity and running time to the methodologies investigated by Yan and Zhang [2022]. In their method, auditing has an additional step: approximating the model through reconstruction before plugging in the estimator. Those methodologies use active learning algorithms for approximating the black-box model i.e, CAL algorithm [Cohn et al., 1994], along with its variant for property active estimation μ-CAL, and its randomized version.

Furthermore, efficient AFA is employed to find significant Fourier coefficients within subsets containing the protected attribute, this model forces search over within subsets containing the protected attribute. In other words, AFA focuses on half of the buckets 2^{n-1} (buckets that contain the protected attribute), where n is the dimension of the input space.

E.3 Additional Experimental Results

Table 4: Estimation error for individual fairness across models and datasets. **Bold** numbers mean lower error.

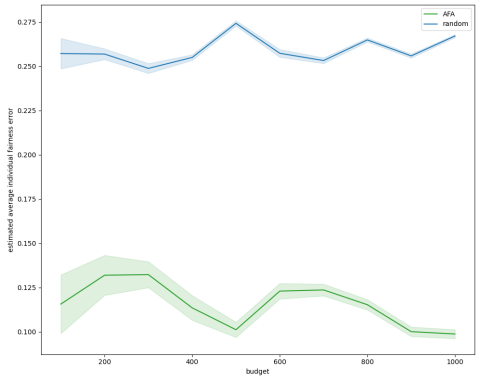
Dataset	COMPAS			Student		
Model	LR	MLP	RF	LR	MLP	RF
`Uniform`	0.050	0.072	0.070	0.12	0.08	0.173
AFA	**0.002**	**0.035**	**0.048**	**0.079**	**0.057**	**0.050**

Individual fairness. For individual fairness, the perturbation parameter l is a free parameter for which Hamming distance measures individual similarity. The parameter l answers the question: *What degree of similarity should the model refrain from distinguishing?* Hence, a good auditor would have the same performance for all possible parameter values l. To evaluate that, we fix $\rho = 0.30$ and compare AFA and random estimator performances for a range of values of parameter l. Experiment details are summarized in Table 5.

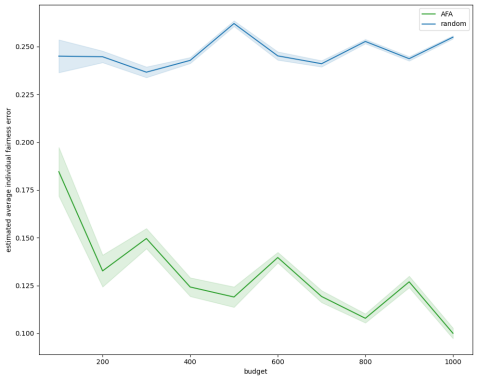
As Figure 4 shows, AFA always outperform random estimator for the property of individual fairness for all different values of perturbation parameter l.

Table 5: A summary of theoretical results: This table summarizes the expression of the estimation for each property with query complexity and computational complexity. **Bold** refers to the best method.

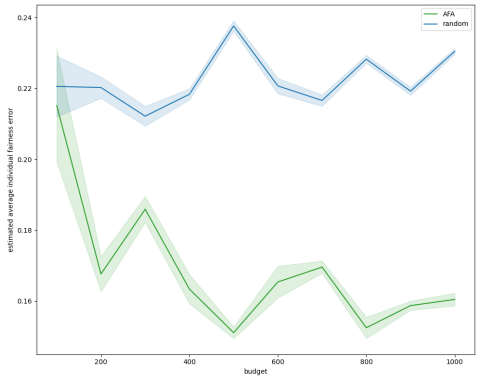
l-parameter	AFA μ_{IFair} error	random μ_{IFair} error
11	**0.123**	0.267
10	**0.119**	0.254
7	**0.141**	0.244
5	**0.169**	0.230
3	**0.166**	0.222



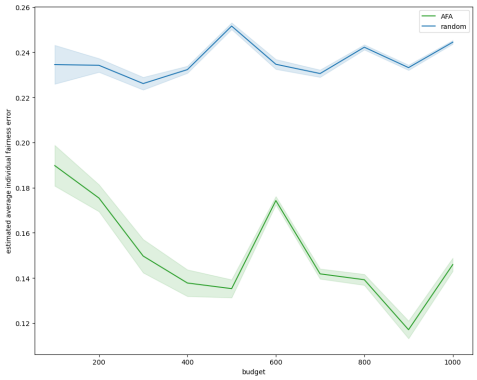
(a) $l = 11$



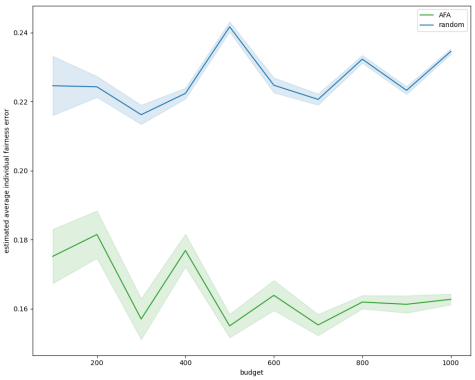
(b) $l = 10$



(c) $l = 7$



(d) $l = 5$



(e) $l = 3$

Figure 4: Comparison of AFA and random estimator on COMPAS dataset for different values of perturbation parameter l.

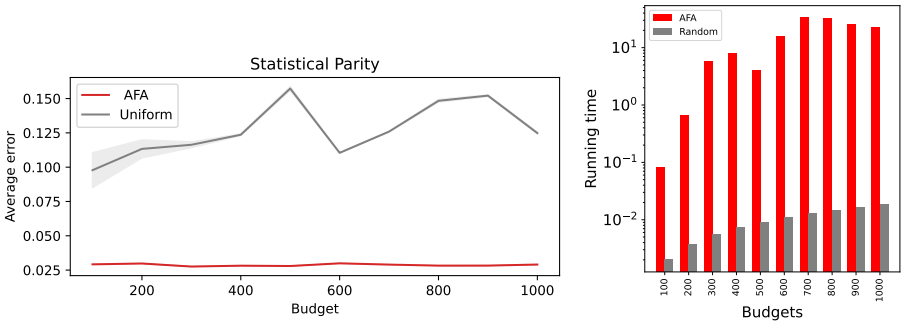


Figure 5: Error (left) and running time (right) of auditors in estimating statistical parity of logistic regression for Student Performance dataset.

Statistical parity. We evaluate SP for the Student Performance dataset, with gender as the protected attribute. Figure 5 shows that AFA's error converges faster to the zero value compared to `Uniform`.

We empirically evaluate the Fourier Pattern for multicalibration by training a logistic regression model on the DRUG dataset, where gender is considered the protected attribute. Figure 6 shows the consistency of AFA performance when the black-box model has multiple outcomes.

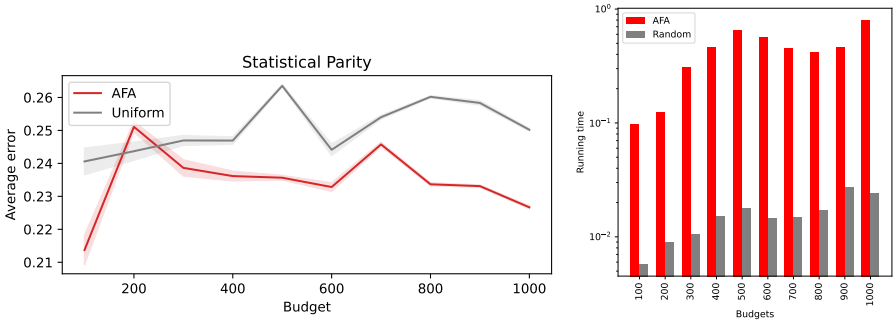


Figure 6: Error (left) and running time (right) of different auditors in estimating statistical parity of logistic regression for Drugs Consumption dataset.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explain clearly the current work done in auditing distributional properties and an extension to verifying those properties.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We emphasize the limitations of the work in the conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the proofs are given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility for experimental results is explained in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to the data and code with sufficient instructions to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting in section 5, we provide details about the black-box model for which we estimate robustness, individual fairness, and group fairness. For group fairness we have multiple baselines derived from the work of previous studies, details about those experiments can be found in the bibliography.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In section 5, we explain statistical significance of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All our computations are performed on an 11th Gen Intel® Core™ i7-1185G7 processor (3.00 GHz, 8 cores) with 32.0 GiB of RAM.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our study focuses on accurately auditing the model's properties. Our proposed model avoids conflicts with societal issues and allows regulatory authorities to maintain accurate estimations of distributional properties. Moreover, it provides firms and companies, whose model is audited, a high degree of freedom to manipulate (manipulation-proof) their decision rules to align with their stakes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts of the work in Appendix **??**.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Data and baselines are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: New assets are well explained theoretically and experimentally.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.