Active Fourier Auditor for Estimating Distributional Properties of ML Models

Ayoub Ajarra ayoub.ajarra@inria.fr
Bishwamittra Ghosh bghosh@u.nus.edu
Debabrota Basu 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.

CONTENTS

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 [Mad21]. These issues are addressed in the growing subfield of ML, i.e. trustworthy or responsible ML [RQG+22, LQL+23], in terms of robustness and fairness of ML models. Robustness quantifies how stable are a model's predictions under perturbation of its inputs [XS11, KNL+20]. Fairness [DHP+12, BHN23] seeks to address discrimination in predictions both at the individual level and across groups. Thus, AI regulations, such as European Union AI Act [Mad21], 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 [RSW+20, WGJ+21, MPR+21, YZ22]. In the context of post-deployment reliability, discovering adversarial examples—inputs incorrectly classified—that are proximate to some training samples is not uncommon [SSRD19, SMB22, GKKW21], as such proximity may naturally occur within the training dataset. This underscores the importance of auditing robustness. Likewise, previous studies [KHL21, TCLD23] have demonstrated that fairness constraint may entail a trade-off with model accuracy. The inherent impossibility of achieving a perfectly fair model further underscores the importance of auditing fairness discrepancies.

Example 1. Following [GBM21, Example 1], let us consider an ML model that predicts who is eligible to get a 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 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 elderly and 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 violates *robustness* when perturbing any feature by an infinitesimal quantity flips the prediction.

Limitations of existing auditing methods. In practice, auditing faces various constraints and challenges. One critical constraint is auditing delay, which can range from days to months due to factors such as access to queries from the dataset catalog and the complexity of the model. This delay prevents companies from manipulating their algorithmic rules to gain competitive advantages. We distinguish between two types of gaming in decisions:

- **Adversarial manipulation**: This type of manipulation occurs at the data point level, where an adversary modifies inputs to the decision rule to maximize some objective function. For example, consider an insurance company that awards points to clients who walk n steps per day. An adversary could artificially replicate the movement associated with walking to gain these points.
- **Manipulation-proof model**: This concept refers to the robustness of a model within a subclass of functions, such that the model remains accurate even under manipulations. Formally, a model is said to be manipulation-proof if, within a specified subclass $\mathcal{H}$, any model $h \in \mathcal{H}$ remains accurate with high probability despite attempts to manipulate it. In our context, similar to [YZ22], we focus on studying this second type of manipulation.

When an external entity (e.g., a company) manipulates[1] the model, it is crucial that the properties of the post-audit model remain closely aligned with the true properties. The concept of manipulation-proof auditors was introduced by [YZ22]. An auditor enjoys manipulation-proof within a subclass of hypotheses if it consistently delivers accurate estimations within that subset.

Formally, for a hypothesis subclass $\mathcal{H}$ and an auditor $\mathcal{A}$, the auditor $\mathcal{A}$ is manipulation-proof if, for any models $h, h' \in \mathcal{H}$ and any manipulation h' designed to exploit the model:

$$\forall h, h' \in \mathcal{H}, \quad \mathbb{P}\left(|\mu(h) - \mu(h')| \leq \epsilon\right) \geq 1 - \delta,$$

where ϵ is a small error tolerance and δ is a small probability of deviation.

Example: The context of manipulation is crucial. Consider a company with a strategy for selecting clients, constrained by fairness towards two protected groups (group A and group B), both of which significantly impact the company's profitability. This strategy, represented by a model h_{pre}, is audited to assess its discriminative bias. Now, suppose the company is developing a new strategy to maximize profit while keeping details confidential to prevent leaks about the new model h_{post}. *Can we ensure that $\mu(h_{pre})$ remains an (ϵ, δ)-PAC estimate of μ?*

In section 3.6, we provide a positive answer to this question.

[1]Here, manipulation refers to modifying the model from its pre-audit to post-audit state.

1.1 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 aare violated above a pre-determined threshold [GNRSY21, JVS20, MS23, HR22, KNRSW18]. But in reality, this threshold might not be known or may vary from one country to the other. Thus, we focus on estimating these properties instead of a 'yes/no' answer, which is a harder problem than verification [GNRSY21]. On estimating distributional properties, [NWE21] proposed a Bayesian approach for estimating properties of black-box optimizers and required a prior distribution of models. [WBH+22] studies simpler distributional properties, e.g. the mean, the median and the treamed mean, using offline and interactive algorithms.[YZ22] 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, which can be a information, and for large models, e.g. deep neural networks. [ADDN17, GBM21] considered finite models for estimating group fairness w.r.t. the features distribution, and [GBM22] further narrowed down to linear models. Therefore, we identify 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 [CRK19, SLR+19], group fairness [ADDN17, GBM21], or individual fairness [JVS20]. (2) **Model-specific auditing:** all the methods considered a prior knowledge about the ML model [NWE21, GBM21, GBM22, YZ22], or a white-box access to the model [CRK19, SLR+19]. 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?*

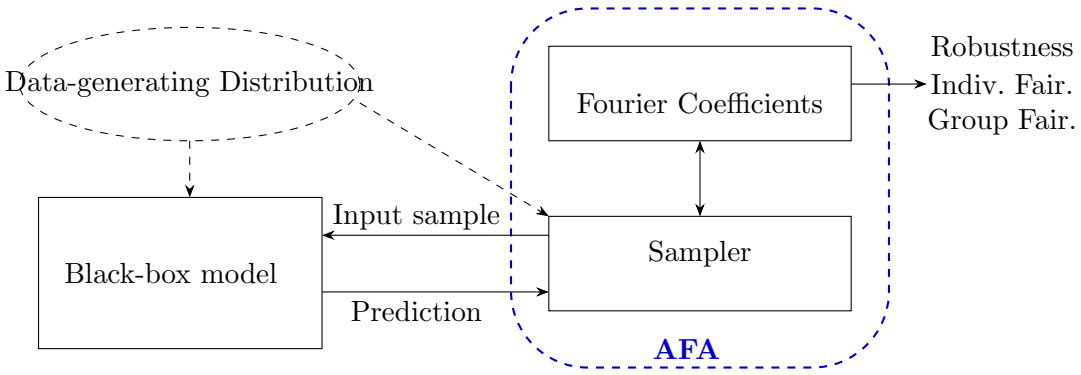

FIG 1. *A schematic of* AFA.

1.2 Contributions.

We propose a framework, namely AFA (**A**ctive **F**ourier **A**uditor), 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 [YZ22]. 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 (Table 1) are summarized as follows.

Properties	**Definition**	**Lower Bound (w.o. MP)**	**Upper Bound (w. MP)**
Robustness	$\mathbb{P}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_\rho(\mathbf{x})}}[h(\mathbf{x}) \neq h(\mathbf{y})]$	✗	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\sqrt{\log(\frac{2}{\delta})}\right)$
Individual fairness	$\mathbb{P}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}}[h(\mathbf{x}) \neq h(\mathbf{y})]$		
Statistical parity	$\mathbb{P}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim\mathcal{D}}}\Big[h(\mathbf{x}) \neq h(\mathbf{y})\Big\vert x_A = 1, y_A = -1\Big]$	$\Omega\left(\frac{8\delta}{3(1-2\alpha)^2\epsilon^2}\right)$	$\mathcal{O}\left(\max\left\{\frac{1}{\epsilon^2}\log\frac{4}{\delta}, \frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right\}\right)$

TABLE 1
Table of properties, definitions, Fourier patterns, lower bounds, and upper bounds.

- *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 [GA89, KM93] 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 $\mathcal{O}\left(\frac{8\sqrt{2}\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right)$ samples to yield (ϵ,δ) estimate of robustness and individual fairness, while it needs $\mathcal{O}\left(\max\left\{\frac{1}{\epsilon^2}\log\frac{4}{\delta}, \frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right\}\right)$ samples to audit group fairness. We provide a lower bound for auditing group fairness that shows we need $\Omega(\frac{8\delta}{3(1-2\alpha)^2\epsilon^2})$ samples. Further, for group fairness, we prove 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: PROPERTIES OF ML MODELS AND FOURIER ANALYSIS

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. Let $n \in \mathbb{N}$. We denote the domain by $\mathcal{X}$, an n-dimensional covariate space, and the codomain by $\mathcal{Y}$. The data-generating distribution over $\mathcal{X} \times \mathcal{Y}$ is denoted by $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$, while the marginal distribution of $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$ over $\mathcal{X}$ is denoted by $\mathcal{D}$. $\mathcal{P}$ denotes a finite pool of unlabeled samples.

Consider a model h, which is a function $h : \mathcal{X} \to \mathcal{Y}$ accessible only through a query oracle (black-box setting). Here, x represents a scalar, and $\mathbf{x}$ represents a vector. $\mathcal{X}$ is a set. We denote $[\![1, n]\!]$ as the set $\{1, \ldots, n\}$.

If $\mathcal{X}$ is a finite space, such as in the case of Boolean features where $\mathcal{X} \triangleq \{-1, 1\}^n$, or categorical features where $\mathcal{X} \triangleq [K]^n$, we denote the power set of $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$.

For all Boolean functions f, g defined on $\mathcal{X}$, we define the scalar product $\langle f, g \rangle_{\mathcal{D}}$, associated to the norm $||f||_{\mathcal{D}} = \sqrt{\langle f, f \rangle_{\mathcal{D}}}$.

2.1 Properties of ML Models

A *distributional property* of an ML model h is defined as a function of the model h, and the underlying distribution $\mathcal{D}$ (aka global): $\mu_h : \mathcal{D} \to \mathbb{R}$. In this paper, we study three such properties, 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 the central quantity for sub-fields of AI, e.g. safe RL [GF15], adversarial ML [KGB16, BR18], and gained attention for safety-critical deployment of AI.

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

Perturbations Γ in the case of evaluating robustness depend on a single degree of freedom ρ. Examples of perturbation mechanisms include Binary feature flipping [O'D14] $\Gamma_\rho(\mathbf{x}) \triangleq \{\mathbf{x}' | \forall i \in [n], \mathbf{x}'_i = \mathbf{x}_i \times \text{Bernoulli}(\rho)\}$, Gaussian perturbation [CRK19] $\Gamma_\rho(\mathbf{x}) \triangleq \{\mathbf{x}' | \mathbf{x}' = \mathbf{x} + \epsilon \text{ and } \epsilon \sim \text{Normal}(0, \rho^2 I)\}$ etc.

In trustworthy and responsible AI, another prevalent concern about deploying ML models is bias in their predictions. This has led to studying different fairness metrics, their auditing algorithms, and algorithms to enhance fairness [MMS+21, BHN23]. There are two types of fairness metrics [BHN23]. The first type is the **individual fairness** that aims to ensure that individuals with similar features should obtain similar predictions [DHP+12].

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

In this context, the neighborhood Γ represents a perturbation mechanism governed by two degrees of freedom: one that encodes the perturbation and another that controls it. The perturbation [2] $\Gamma(\mathbf{x})$ is commonly defined as the points around $\mathbf{x}$ that are at a distance less than $\rho \geq 0$ w.r.t. a pre-defined distance metric. The distance metric depends on the application of choice and the input data [MMS+21].

[2] For simplicity, we use the same notation for both the perturbation mechanism and the neighborhood.

Group fairness is the other type 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 [MMS+21]. Specifically, we focus on *Statistical Parity (SP)* [FFM+15, DHP+12] 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). Let us consider that each input $\mathbf{x}$ consists of a component representing a sensitive group $x_A \in \{-1, 1\}$. The statistical parity of h is:

$$\mu_{\mathrm{GF}}(h) \triangleq \max_{y \in \mathcal{Y}} |\, \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x}) = y | x_A = 1] - \mathbb{P}_{\mathbf{x}' \sim \mathcal{D}}[h(\mathbf{x}') = y | x_A = -1]|$$

In AFA, we use techniques of Fourier analysis to design one computational scheme for simultaneously estimating these three properties of an ML model. This scheme can be generalized to any other distributional property.

2.2 A Primer on Fourier Analysis

Designing AFA is motivated by Boolean function analysis, which allows the decomposition of a function on a basis dependent on the input space dimension [O'D14]. Specifically, Theorem 1 decomposes any bounded-output model h on the Boolean input domain w.r.t. the basis of parity functions.

Theorem 1 ([dW08]). *If the distribution over $\mathcal{X}$ is uniform, i.e. $\mathcal{D} = \mathcal{U}(\mathcal{X})$, any bounded function $h : \mathcal{X} \to \mathcal{Y}$ can be uniquely expressed as $h(\mathbf{x}) = \sum_{S \subseteq [\![1,n]\!]} \hat{h}(S)\chi_S(\mathbf{x})$, where for all $S \subseteq [\![1, n]\!]$, $\hat{h}(S) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{U}([\![1,n]\!])}[h(\mathbf{x})\chi_S(\mathbf{x})]$. Here, $\chi_S(\mathbf{x}) \triangleq \prod_{i \in S} x_i$ for $S \in \mathcal{P}(\mathcal{X})$.*

If S is a subset of $[\![1, n]\!]$, and $\mathbf{x} \sim \mathcal{D}$, we denote by $\chi_S(\mathbf{x})$ the monomial defined as the product of x's coordinates belonging to S. Since the set of parity functions encompasses all the monomials $\{\chi_S\}$ indexed by $S \in [\![1, n]\!]$, *any classifier with binary input and finite number of labels can be expressed with the basis of parity functions*, if the data-generation distribution is uniform.

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 $\{-1, 1\}^2$.

However, in real-world auditing tasks, the distribution is unknown, and non-uniform. Theorem 1 is further extended to the distribution agnostic setting. Gram-Schmidt orthogonalization generalizes the Fourier expansion w.r.t. any *unknown* distribution over the input space.

Proposition 1 ([HSSS21]). *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*

$$h(\mathbf{x}) = \sum\nolimits_{S \subseteq [n]} \hat{h}(S)\psi_S(\mathbf{x}) \quad \text{for any} \quad x \sim \mathcal{D}. \tag{2.1}$$

Additionally, the Fourier coefficients $\hat{h}(S) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x})\psi_S(\mathbf{x})]$ are unique for all $S \in \mathcal{P}(\mathcal{X})$.

Since we aim to follow the agnostic PAC setting for auditing an unknown model with unknown data-generating distribution, we consider an adaption of Proposition 1. Here, the predicted label can depend on the input features and some underlying randomness generating the labels. This distribution agnostic is not a new result. For completeness, we provide a proof in Appendix A.

Influence functions. To estimate the properties of interest, we use a tool from Fourier analysis, i.e. *influence functions* [O'D14]. Influence functions measure how changing an input changes the output of a model. Influence functions of different forms are widely used in statistics, e.g. to design robust estimators [MBM22], and ML, e.g. to find important features [HSSS21], to evaluate how features induce bias [GBM21], to explain contribution of datapoints on predictions [IMPE+22]. 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 $\mathrm{Inf}_\Gamma(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \neq h(\Gamma(\mathbf{x}))]$. We call $\mathrm{Inf}_\Gamma(h)$ to be *deterministic* (resp. *randomised*) if the transformation Γ is deterministic (resp. randomized).

In general, deterministic influence functions are used in Boolean function analysis [O'D14]. In contrast, in Section 3, we express robustness, individual fairness using randomized influence functions, while for group fairness we express it with a randomized version of the deterministic influence function [O'D14]. We also show that the influence functions can be computed using the Fourier coefficients of the model under audit (Equation (2.1)).

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 [**?**], the auditor is denoted as $\hat{\mu} : \mathcal{H} \times \mathcal{P} \to \mathbb{R}$. Here, the auditor makes use of a pool of unlabeled samples $\mathcal{P}$ and applies active learning techniques (e.g. CAL algorithm) to query samples. This process uses the additional information provided by the hypothesis class where the model h lives. Following the reconstruction phase, the auditor has an approximate model $\hat{h}$ of the true model h, enabling the 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 auditing algorithms, AFA uses Fourier expansion and adaptive queries to directly estimate the robustness, Individual Fairness (IF), and Group Fairness (GF) properties of a model h. In this scenario, the auditor is defined as $\hat{\mu} : \mathcal{F}_\mu \times \mathcal{P} \to \mathbb{R}$, 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 use an adaptive sampling to compute the Fourier coefficients and thus, estimate the properties at once (Algorithm 1).

We begin by defining a PAC-agnostic auditor that we try and 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 $m : (0,1)^2 \to \mathbb{N}$ such that for all $m \geq m(\epsilon,\delta)$ 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 $\mu(h)$ [KNRSW18]. Any distributional property, including robustness, individual fairness and group fairness, is computable given the existence of the uniform estimator.

3.1 The extra cost of reconstruction.

The most trivial way to estimate a model property is to reconstruct the model and then use a plug-in estimator [YZ22]. 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 [JCB+20].

PROPOSITION 2 (Cost of Reconstruction). *If $\hat{h}$ is the reconstructed model from h, then:*

$$|\mu_{\text{GFair}}(\hat{h}) - \mu_{\text{GFair}}(h)| \leq \min\left\{1, \frac{\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x}) \neq h(\mathbf{x})]}{\min(\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = 1], \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}_A = -1])}\right\}. \tag{3.1}$$

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 in Appendix B. This motivates an approach that avoids model reconstruction by computing only the right components of the model expansion.

In order to capture the information relevant for estimating our properties of interest, we will represent them in terms of Fourier coefficients given in the model decomposition. In contrast, model reconstruction requires recovering all the Fourier coefficients in the decomposition on the Gram-Schmidt basis. In this section, we compute our properties of interest: robustness, individual fairness and group fairness in terms of deterministic and random influence functions. Those functions can be expressed in terms of the Fourier coefficients of the model h, allowing a representation of properties to audit in terms of Fourier coefficients. The problem reduces to an estimation of Fourier coefficients.

3.2 Fixing the basis.

In the following proposition, we construct a distribution-independent basis for Fourier expansion. This basis depends on the perturbation protocol Γ. The protocol defines robustness and individual fairness, allowing their computation in terms of the Fourier coefficients.

PROPOSITION 3. *There exists an orthonormal basis $\{\psi_S\}_{S\in\mathcal{P}(\mathcal{X})}$ for classifiers that verifies the following:*

- *Homogeneity:*[3] $\forall \mathbf{x} \in \mathcal{X}, \underset{\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}{\mathbb{E}}[\psi_S(\mathbf{y})] = \rho^{|S|}\psi_S(\mathbf{x})$
- *Stochastic anisotropy:*[4] $\forall \mathbf{x} \in \mathcal{X}, \underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}[\psi_S(\mathbf{y})] = \frac{\rho^{|S|}}{n^l}\psi_S(\mathbf{x})$

The proof is deferred in Appendix C.

[3]For robustness.
[4]for individual fairness.

Remark. The properties that the basis satisfies allow us to find the Fourier pattern for robustness and individual fairness properties. Since group fairness does not depend on a perturbation protocol encoded in the basis construction, we can observe that group fairness property doesn't depend on the basis.

For the rest of this paper, we fix our basis to be the one constructed in Proposition 3.

3.3 Model properties through Fourier Expansion.

Now, we express the three model properties of h using its Fourier coefficients.

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 rho-flipping influence function.

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

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 the Fourier coefficients of h.

Proposition 4. *Robustness of h under the Γ_ρ flipping perturbation is equivalent to the ρ-flipping influence function, and thus, can be expressed as*

$$\mu_{\mathrm{Rob}}(h) = \mathrm{Inf}_\rho(h) = \sum_{S\subseteq[n]} \rho^{|S|}\hat{h}(S)^2. \tag{3.2}$$

The proof is deferred to Appendix D.1.

b. Individual Fairness (IF). To demonstrate the universality of our approach, we express IF using the model's Fourier coefficients. IF of a model measures its capacity to yield similar predictions for similar input features of individuals [DHP+12, AFSV16]. The similarity between individuals are depicted through different distance 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 $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$* [AFSV16].

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 neighbourhood around each input sample. Specifically, we consider the perturbation mechanism $\Gamma_{\rho,l}(\cdot)$ 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. Pairs $(\mathbf{x}, \mathbf{x}')$

are referred to as (ρ, l)-correlated pairs. 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 $\mathrm{Inf}_{\rho,l}(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D} \\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}} [h(\mathbf{x}) \neq h(\mathbf{y})]$.

We leverage (ρ, l)-flipping influence function to express IF of h in terms of its Fourier coefficients (Proposition 5). The proof is in Appendix D.2.

PROPOSITION 5. *Individual fairness under the $\Gamma_{\rho,l}$ perturbation is equivalent to the (ρ, l)-flipping influence function, and thus can be expressed as:*

$$\mu_{\mathrm{IFair}}(h) = \mathrm{Inf}_{\rho,l}(h) = \sum_{S\subseteq[n]} \frac{\rho^{|S|}}{n^l} \hat{h}(S)^2. \tag{3.3}$$

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 (3.2) and (3.3), we observe that both the properties as $\mu_\cdot(h) = \sum_{S\subseteq[n]} \mathtt{char}(S, \mu_\cdot)\hat{h}(S)^2$, such that $\mathtt{char}(S, \mu_{\mathrm{Rob}}) = \rho^{|S|}$, and $\mathtt{char}(S, \mu_{\mathrm{IFair}}) = \frac{\rho^{|S|}}{n^l}$. We call $\mathtt{char}$ 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 groups of populations in the data [BHN23]. We specifically focus on Statistical Parity (SP) as the measure of deviation from GF [DHP+12, FFM+15]. 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 $\mathrm{Inf}_A(h) \triangleq \mathop{\mathbb{P}}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}} \Big[h(\mathbf{x}) \neq h(\mathbf{y}) \Big| x_A = 1, y_A = -1\Big]$.

$\mathrm{Inf}_A(h)$ expresses the probability that the outcome of h changes when changing group membership within the generating distribution $\mathcal{D}$. 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'D14]. 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 6. *Statistical parity of h w.r.t a sensitive attribute A and distribution $\mathcal{D}$ is the root of the second order polynomial:*

$$P(X) \triangleq 2\alpha(1-\alpha)X^2 - \hat{h}(\emptyset)(1-2\alpha)X - \sum_{\substack{S\subseteq[n] \\ S\ni A}} \hat{h}(S)^2 - \frac{(1-\hat{h}^2(\emptyset))}{2} \tag{3.4}$$

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

Additionally, if $\mathcal{D}$ is the uniform distribution, the Fourier coefficient of this sensitive attribute fully captures statistical parity.

COROLLARY 2. *If $\mathcal{D}$ is the uniform distribution on $\mathcal{X}$, $\mu_{\text{GFair}}(h) = \hat{h}(\{A\})$.*

The proofs are deferred to Appendix D.

Summary of the Fourier patterns 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 $\hat{h}(\emptyset)$.

3.4 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 $\hat{h}(S) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\psi_S(\mathbf{x})]$, 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 show that even if we want to exactly compute only the significant Fourier coefficients of h with values larger than a threshold τ, the problem is NP-complete.

THEOREM 3. *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.*

This 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.

3.5 Algorithm: Active Fourier Auditor (AFA)

In the previous section, we have shown that finding significant Fourier coefficients can be an NP-hard problem. In this section, we propose algorithm 1, that takes as input *restricted access* of $q > 0$ queries and request labels from the black-box oracle of h. Those queries enable us to find the squares of significant Fourier coefficients and estimate them simultaneously. The list of significant Fourier coefficients L_h [5] of the model's h contains both subsets and their estimated Fourier weights. Since the properties – robustness, individual fairness and group fairness – depend on those Fourier coefficients, we plug in their estimates and output an (ϵ, δ)-PAC estimate of our properties.

Algorithmic insights: To find subsets of significant Fourier coefficients, we start with the power set. By Parseval's identity, we know that the weight of the power set is 1, that is:

$$\sum_{S\in\mathcal{P}(\mathcal{X})} \hat{h}(S)^2 = 1$$

[5] L_h is a random variable.

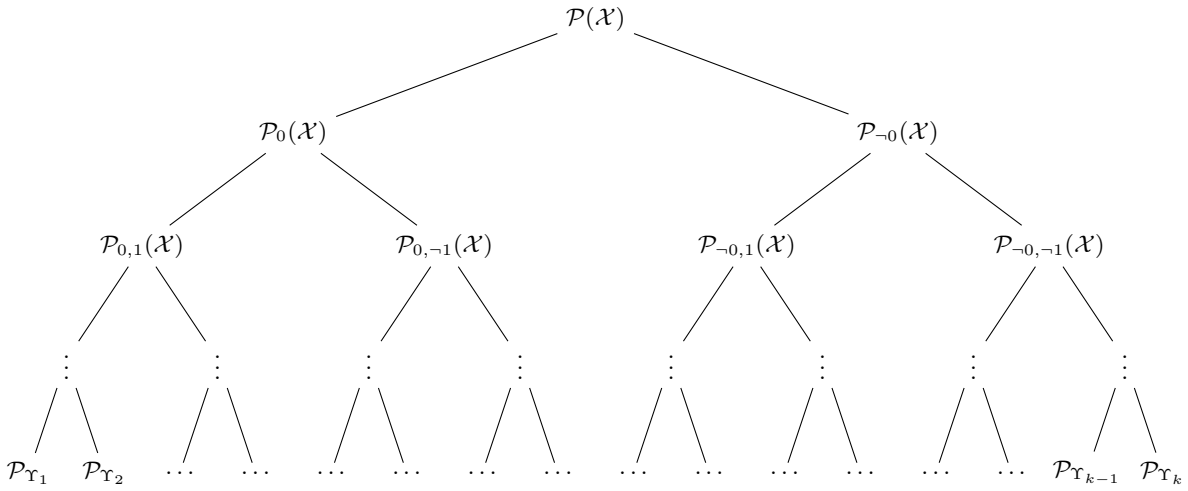

FIG 2. *The algorithm begins with the set of all Fourier coefficients, their weight is* 1*, this weight is above the threshold which is generally small (Boolean functions). The algorithm proceeds by splitting the bucket and verifies at each level of the tree the weight of the node: If the weight falls below the threshold, the algorithm halts. Otherwise, the algorithm continues to expand, defining a set of (informative) trajectories* Υ*, the subsets with large Fourier coefficients are* $\{\mathcal{P}_{\Upsilon_1}(\mathcal{X}), \cdots, \mathcal{P}_{\Upsilon_k}(\mathcal{X})\}$.

Let Υ denote a trajectory starting from the set of all Fourier coefficients. *From* Υ *(the power set), how can we reach subsets of Fourier coefficients that are above a given threshold* τ*?*

We denote subsets that contain element i as $\mathcal{P}_i(\mathcal{X})$, and subsets that do not contain element i as $\mathcal{P}_{\neg i}(\mathcal{X})$.

That is:

$$\mathcal{P}_i(\mathcal{X}) = \{S \in \mathcal{P}(\mathcal{X}) : i \in S\}$$
$$\mathcal{P}_{\neg i}(\mathcal{X}) = \{S \in \mathcal{P}(\mathcal{X}) : i \notin S\}$$

A formal language Υ is defined as a sequence of words $\{(i, \neg i)\}_{i \in [\![1,n]\!]}$. The language Υ encodes the trajectory to significant Fourier coefficients. We use the Goldreich-Levin algorithm [GA89, KM93] to find the trajectory to large Fourier coefficients as shown in Figure 3.5.

Illustrative example: To illustrate how AFA algorithm works, we consider the simple case where $n = 3$, the domain is $\mathcal{X} = \{-1, 1\}^3$, the features of the domain $\mathcal{X}$ are $\{0, 1, 2\}$. The number of possible feature combinations is 8, given by the number of subsets of $\{0, 1, 2\}$, that we denote $\mathcal{P}(\mathcal{X})$. Figure 3 illustrates how AFA proceeds:

Let $k \in [\![1, n]\!]$ and $S \subseteq [k]$. We define the buckets as follows:

$$\mathcal{B}^{S,k} \triangleq \Big\{ S \cup T \mid T \subseteq \{k+1, \ldots, n\} \Big\},$$

and their corresponding Fourier weights:

$$\mathcal{W}^{S,k} \triangleq \sum_{T \subseteq \{k+1,\ldots,n\}} \hat{f}(S \cup T)^2.$$

The Goldreich-Levin algorithm is a recursive algorithm that searches for significant Fourier coefficients by estimating the weights $\mathcal{W}^{S,k}$. Goldreich-Levin operates as follows: First, the bucket is

initialized at $\mathcal{B}^{\emptyset,0}$, which represents the weight over all subsets of $[\![1,n]\!]$; this weight is equal to 1 by Parseval's identity. The bucket $\mathcal{B}^{S,k}$ is then split into two buckets of the same cardinality: $\mathcal{B}^{S,k-1}$ and $\mathcal{B}^{S\cup\{k+1\},k+1}$. Goldreich-Levin then estimates the weight of each bucket by black-box queries to the model to audit h, where the query is made by fixing z (the last $n-k$ elements of the vector) and randomly sampling over the first k elements. This allows computation of the weight of the bucket via the Fourier transform of the function h restricted to the first k directions. The algorithm discards the bucket whose weight is below the threshold τ. When all collected buckets at round t have exactly one element (bucket of unique subset) corresponding to $k=n$ (the depth of the tree), the algorithm halts, as the buckets collected so far are exactly subsets of $[\![1,n]\!]$ with large Fourier coefficients.

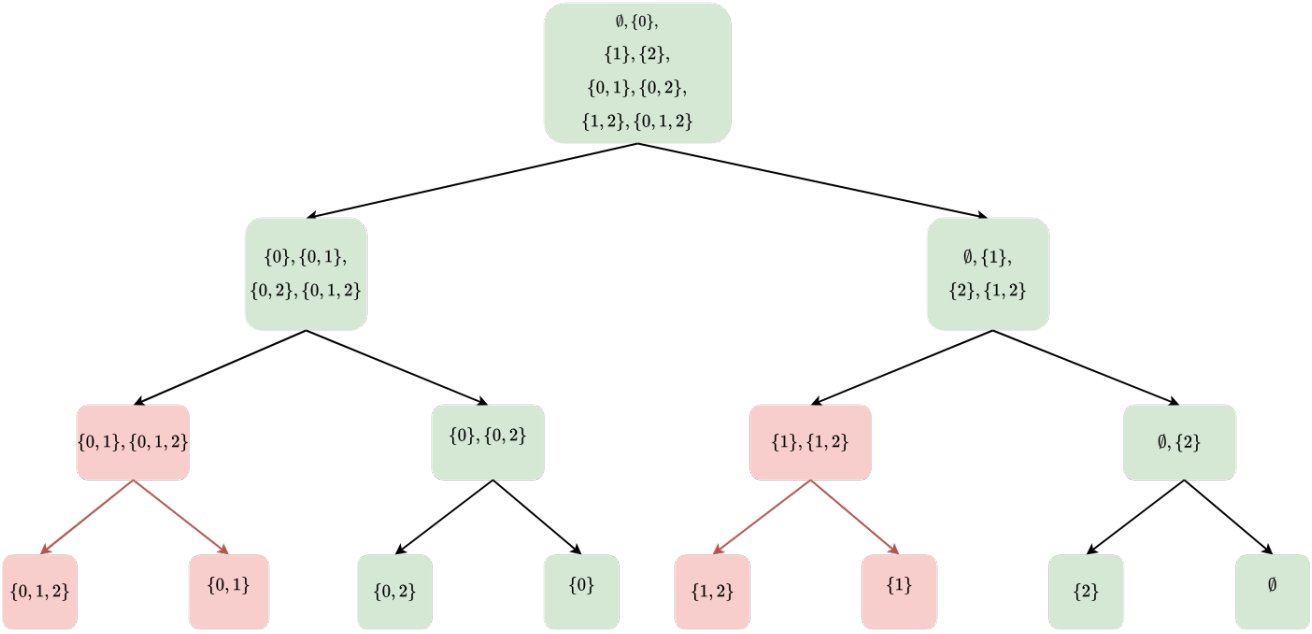

FIG 3. *The algorithm begins with the power set, and compute the weights of buckets at each level of the tree. In this example the weight of the buckets* $\{\{0,1\},\{0,1,2\}\}$ *and* $\{\{1\},\{1,2\}\}$ *do not satisfy the threshold constraint, therefore* AFA *halts on those buckets (nodes) and proceeds with remaining nodes.*

Algorithm 1 Active Fourier Auditor (AFA)

Input: Sensitive attribute A, Query access to h, $\tau, \delta \in (0,1)$, $\epsilon \leftarrow \tau^2/4$
$\{x_k, h(x_k)\}_{k\in[q]} \leftarrow \mathrm{MQ}(h,q)$
$L_h \leftarrow \textsc{GoldreichLevin}(h,q,\tau,\delta)$
$\hat{\mu}(h) \leftarrow \sum_{S\in L_h} \mathtt{char}(\mu,S)\hat{h}(s)^2$
$\hat{\mu}_{GF}(h) \leftarrow c\Big(\Big(1-\frac{1}{c^2}+\frac{\mathrm{Inf}_A(h)}{2\alpha(1-\alpha)c^2}\Big)^{1/2}-1\Big)$
return $\{\hat{\mu}_{RB}, \hat{\mu}_{IF}, \hat{\mu}_{GF}\}$

Extension to categorical domains: For simplicity, we explain our methodology only for the Boolean input domain. However, AFA is equally applicable to the categorical domain, where the new basis changes from parity functions to unity root functions. Thus, the new Fourier expansion

becomes

$$h(\mathbf{x}) = \sum_{\zeta} \hat{h}(\zeta)\omega_p(\langle \zeta, \mathbf{x}\rangle).$$

We express our properties in terms of Fourier coefficients on this basis:

$$\mu_{\text{Rob}}(h) = \left(\frac{1}{1-\rho}\right)^n \sum_{\zeta:\zeta_j\neq 0} |\hat{h}(\zeta)|^2 + \sum_{\zeta:\zeta_j=0} |\hat{h}(\zeta)|^2,$$

$$\mu_{\text{IFair}}(h) = \frac{1}{p}\sum_{\zeta} |\hat{h}(\zeta)|^2 \cos\left(\frac{2\pi l}{p(1+\rho)} \sum_{\mathcal{T}\in\mathbb{F}_p^n} \langle \zeta, \mathcal{T}\rangle\right),$$

$$\mu_{\text{GFair}}(h) = \frac{2p}{p-1}\sum_{\zeta} |\text{supp}(\zeta)||\hat{h}(\zeta)|^2.$$

We further extend the computational hardness result to the categorical case. Appendix F provides an analysis of these results.

In the next section, we provide upper bounds on the sample complexity of AFA. Further, for group fairness, we prove AFA is manipulation-proof under perturbation of 2^{n-1} Fourier coefficients.

3.6 Upper bounds in manipulation-proof auditing

Rethinking manipulation-proof definition. Despite adopting the manipulation-proof framework similar to [YZ22], their approach does not adequately capture the relevant information about the property of interest (SP). Their method primarily revolves around fully reconstructing the model, defining 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 concerning each protected group (hard functions). In contrast, our proposed definition captures those hard functions by defining the essential information required for auditing statistical parity (Fourier coefficients).

Answering this question requires the construction of a manipulation proofness region in the pre-audit phase, which we denote $\mathcal{M}$.

Definition 9 (Fourier strategic manipulation-proof). Let h be a model that admits a Fourier expansion as in equation 2.1:

$$h = \sum\nolimits_{S\subseteq[n]} \hat{h}(S)\psi_S$$

We say that an auditor $\mathcal{A}$ achieves optimal manipulation-proofness 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:

$$\forall h, h' \in \mathcal{M}, \mathbb{P}\left(|\mu(h) - \mu(h')| \leq \epsilon\right) \geq 1-\delta, \tag{3.5}$$

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

THEOREM 5 (Upper bounds for estimating robustness and individual fairness). AFA *is a PAC-agnostic auditor for robustness and individual fairness with a sample complexity* $\mathcal{O}\left(\frac{8\sqrt{2}\,\texttt{char}(L,\mu)(1-4\,\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right)$ *Here,* $\texttt{char}(L,\mu.) \triangleq \sum_{S\in L}\texttt{char}(S,\mu.)$, $\texttt{char}(\bar{L},\mu.) \triangleq \sum_{S\in\bar{L}}\texttt{char}(S,\mu.)$.

THEOREM 6 (Upper bounds for GF). AFA *yields an* (ϵ,δ)*-PAC estimate of* $\mu_{\text{GFair}}(h)$ *if it has access to predictions of* $\mathcal{O}\left(\max\left\{\frac{1}{\epsilon^2}\log\frac{4}{\delta}, \frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right\}\right)$ *input samples.*

The proofs of these theorems can be found in Appendix E.

Specifically, we demonstrate that AFA achieves an optimal rate of $\frac{1}{\epsilon^2}$ for robustness and individual fairness and an $\frac{1}{\epsilon^4}$ 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.

4. LOWER BOUNDS IN THE ABSENCE OF MANIPULATION-PROOF

In the following, we propose a lower bound on 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 **??**.

THEOREM 7 (Lower bounds). *Let* $\epsilon \in (0, 7/\sqrt{d}]$, $\delta \in (0, 1/2]$. *We aim to obtain* (ϵ,δ)*-PAC estimate of SP of model* $h \in \mathcal{H}$*, such that 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 when* $\mathcal{A}$ *is given* $\Omega(\frac{8\delta}{3(1-2\alpha)^2\epsilon^2})$ *samples, it outputs an estimate* $\hat{\mu}$ *of* $\mu_{\text{GFair}}(h^*)$*: with large probability error:* $\mathbb{P}[|\hat{\mu} - \mu_{\text{GFair}}(h^*)| > \epsilon] > \delta$.

Theorem 7 shows that estimating statistical parity without reconstruction requires $\Omega(\frac{8}{3(1-2\alpha)^2\epsilon^2})$ samples. This result extends the existing sample complexity results with model reconstruction [YZ22], and also provides a reference of optimality for upper bounds. We highlight the gap from the upper bound established in Theorem 6, attributed to the lack of the manipulation-proof.

5. EMPIRICAL PERFORMANCE ANALYSIS

In this section, we evaluate the performance of AFA to estimate robustness, individual fairness, and statistical parity of multiple models. Below we discuss experimental setup, objectives, and results.

Experimental Setup. We conduct experiments on COMPAS dataset [ALMK16], where the prediction task is to decide whether a person will re-offend crimes in the next two years given the demographic of the person. We consider three ML models, namely Random Forest (RF), Logistic Regression (LR) and Multi-layer perceptron (MLP). For robustness, we consider the feature flipping perturbation with a flipping probability $\rho \in [0, 1/2]$. For individual fairness, we vary the flipping probability $\rho \in [0, 1/2]$ and set number of features under flip as $l = 10$ where $n = 13$. For robustness and individual fairness, we compare AFA with a uniform random sampling method, namely `Uniform`. For group fairness, we compare AFA with `Uniform` and CAL-based active fairness auditor [YZ22,

TABLE 2
Average estimation error for statistical parity across different ML models. '—' denotes when a method cannot scale to the model.

Model	μCAL	CAL	Randomized μCAL	Uniform	Inefficient AFA	AFA
Logistic Regression	0.315	0.315	0.312	0.077	0.012	**0.006**
MLP	—	—	—	0.225	0.149	**0.147**
Random Forest	—	—	—	0.077	0.012	**0.006**

Algorithm 3], namely CAL and its variants μCAL and randomized μCAL. Each experiment is run 10 times and we report the average results. We refer to the Appendix G.1 for details.

Our empirical studies have the following objectives:

1. How accurate AFA is w.r.t. Uniform to audit robustness and individual fairness for different ρ?
2. How accurate, sample-efficient, and scalable AFA is with baselines in estimating statistical parity across different ML models?

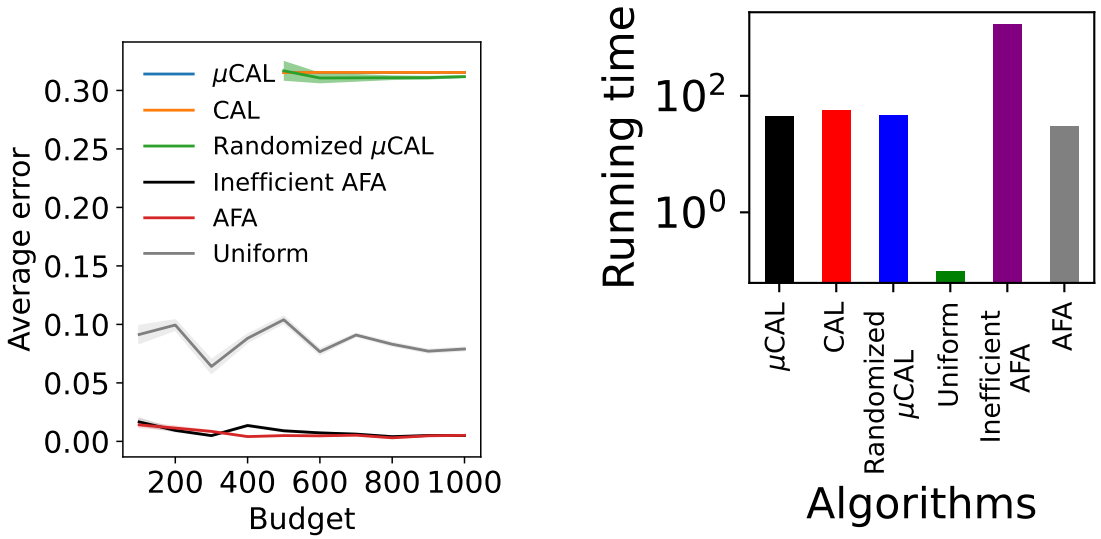

FIG 4. *Error (left) and running time (right) of different auditors in estimating statistical parity.*

Summary of Results. AFA achieves lower estimation error compared to Uniform while estimating robustness and individual fairness across perturbation levels. Compared to CAL and its variants in estimating group fairness, AFA not only achieves lower estimation error, but also incurs lower computation time across models and number of samples. Details are the following.

Accurate Estimation of Robustness and Individual Fairness. Table 3 demonstrates the estimation error of AFA and Uniform for different ρs and 1000 samples. The ground-truth of the property is estimated on the dataset. In each value of ρ, AFA achieves the lowest estimation error than Uniform. *Therefore,* AFA *is more accurate than* Uniform *in estimating robustness and individual fairness.*

Accurate, Sample Efficient, and Fast Estimation of Group Fairness. Table 2 demonstrates the estimation error of group fairness by different methods. In LR model, CAL incurs higher error than Uniform and AFA and cannot scale on MLP and RF. In contrast, AFA *appears as the*

most accurate method by incurring the lowest error across different ML models.

Figure 4 (left) demonstrates the sample efficiency of different methods for statistical parity. AFA requires the lowest number of samples to reach almost zero estimation error – thereby, AFA *is sample efficient than other methods.* Figure 4 (right) demonstrates the runtime for estimating statistical parity, where AFA is the second fastest method after `Uniform` and faster than `CAL`. *Therefore,* AFA *is accurate, sample efficient, and fast compared to baselines in estimating group fairness.*

6. CONCLUSION AND FUTURE WORKS.

In this study, we introduced a model-agnostic and black-box method for auditing the properties of machine learning models. We focused on three properties: robustness, individual fairness, and group fairness. By establishing the intrinsic connection between these properties and the Fourier coefficients of our black-box model, we demonstrate that estimating significant Fourier coefficients enables us to achieve a PAC approximation of all these properties simultaneously. Our method, AFA, is specifically designed for computing these properties efficiently.

Numerical experiments validate that AFA outperforms baseline algorithms in accuracy and scalability across various models and parameter configurations, affirming its universality and efficiency. Currently, AFA applies only to binary and categorical domains.

Looking ahead, our interest lies in extending AFA-like methodologies to models with continuous input domains. However, a significant challenge lies in identifying a suitable basis for expanding functions with non-categorical inputs.

7. ACKNOWLEDGEMENTS

This work was supported by the Regalia Project Grant.

REFERENCES

[ADDN17] 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.

[AFSV16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. In *arxiv*, 2016. arxiv:1609.07236.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[BHN23] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[BR18] 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.

[CAR94] David Cohn, Les Atlas, and Ladner Richard. Improving generalization with active learning. In *Machine Learning (20)*, pages 201–221, 1994.

[CRK19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[DHP+12] 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.

[dW08] Ronald de Wolf. *A Brief Introduction to Fourier Analysis on the Boolean Cube*. Theory of Computing Library, 2008.

[FFM+15] 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.

[GA89] 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.

[GBM21] 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.

[GBM22] 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.

[GF15] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[GKKW21] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. In *Journal of Machine Learning Research (JMLR)*, volume 22, pages 1—29, 2021.

[GNRSY21] Shafi Goldwasser, Guy N. Rothblum, Jonathan Shafer, and Amir Yehudayoff. Interactive proofs for verifying machine learning. *Innovations in Theoretical Computer Science Conference (ITCS)*, 2021.

[HR22] 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.

[HSSS21] 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.

[IMPE+22] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

[JCB+20] 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.

[JVS20] 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.

[KGB16] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2016.

[KHL21] Nikola Konstantinov and Christoph H. Lampert. On the impossibility of fairness-aware learning from corrupted data. In *Algorithmic Fairness through the Lens of Causality and Robustness*, 2021.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. *SIAM J. Comput., 22(6):1331–1348*, 1993.

[KNL+20] 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.

[KNRSW18] 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.

[LQL+23] 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.

[Mad21] Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

[MBM22] 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.

[MMS+21] 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.

[MPR+21] 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.

[MS23] Saachi Mutreja and Jonathan Shafer. Pac verification of statistical algorithms. *36th Annual Conference on Learning Theory (COLT)*, 2023.

[NWE21] 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.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, Massachusetts, 2014.

[RQG+22] 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.

[RSW+20] 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.

[SLR+19] 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.

[SMB22] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. In *arxiv*, 2022. arXiv:2106.10151.

[SSRD19] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. In *arxiv*, 2019.

arXiv:1901.10861.

[TCLD23] Hua Tang, Lu Cheng, Ninghao Liu, and Mengnan Du. A theoretical approach to characterize the accuracy-fairness trade-off pareto frontier. In *arxiv*, 2023. arXiv:2310.12785.

[WBH+22] 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.

[WGJ+21] 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.

[XS11] Huan Xu and Mannor Shie. Robustness and generalization. In *Maéch Learn*, volume 86, pages 391—423, 2011.

[YZ22] Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pages 24929–24962. PMLR, 2022.

APPENDIX A: GRAM-SCHMIDT ORTHOGONALIZATION FOR FOURIER DECOMPOSITION

Proof of proposition 1: 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

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

Additionally, the Fourier coefficients $\hat{h}(S) \triangleq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\psi_S(\mathbf{x})]$ are unique for all $S \in \mathcal{P}(\mathcal{X})$.

Gram-Schmidt process enriches the Fourier structure of Boolean function, as given in the following proof [HSSS21]:

Proof. Let

$$X = (X_1, \ldots, X_n) \sim \mathcal{D}$$

For $i \in [\![1, n]\!]$, let μ_i (resp. σ_i) denote the mean (resp. variance) of the random variable X_i. We consider the $S-$concentrated functions, which is a centered and normalized version of the parity functions, defined as:

$$\forall S \in \mathcal{P}(\mathcal{X}), \forall x \in \mathcal{X}, \psi_S(x) \triangleq \prod_{i\in S} \frac{x_i - \mu_i}{\sigma_i}$$

This set of functions is not orthogonal without the assumption of correlated features. Thus we derive an orthogonal family of functions based on the Gram-Schmidt process.

Consider the following representation of power sets in this order:

$$\varnothing, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \ldots, \{1,2,3,\ldots,n\}$$

Starting from the set of functions $\{\chi_S\}$, we construct the new distribution-free basis for Fourier expansion using the Gram-Schmidt orthogonalization process, that is:

$$\psi_\emptyset = 1$$

$$\forall i \in [\![1, 2^n]\!] : \psi_{S_i} = \chi_{S_i} - \sum_{j=1}^{i-1} \langle\, \psi_{S_j}, \chi_{S_i}\rangle_{\mathcal{D}} \psi_{S_j}$$

$$\psi_{S_i} = \begin{cases} \frac{\tilde{\psi}_{S_i}}{\|\tilde{\psi}_{S_i}\|_{2,\mathcal{D}}} & \text{if } \tilde{\psi}_{S_i} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

One can easily verify by construction that $\{\psi_{S_i}\}_{i\in[2^n]}$ is indeed a basis for Boolean functionals, which concludes the proof. □

APPENDIX B: THE COST OF AUDITING WITH RECONSTRUCTION

Proof of proposition 2: If $\hat{h}$ is the reconstructed model from h, then:

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

PROOF.

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

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

On the other hand

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

By symmetry of h and $\hat{h}$ roles

$$\left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x})=1|x_A=1]-\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})=1|x_A=1]\right|\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x})\neq h(\mathbf{x})|x_A=1] \tag{B.3}$$

Similarly:

$$\left|\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x})=1|x_A=0]-\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})=1|x_A=0]\right|\leq \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[\hat{h}(\mathbf{x})\neq h(\mathbf{x})|x_A=0] \tag{B.4}$$

On the other hand,

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

Where the first step comes from the triangle inequality, the second step comes from inequalities B.3 and B.4, and the last step comes from inequality B.2. □

APPENDIX C: DISTRIBUTION-INDEPENDENT BASIS CONSTRUCTION.

Proof of proposition 3:

There exists an orthonormal basis $\{\psi_S\}_{S\in\mathcal{P}(\mathcal{X})}$ for classifiers that verifies the following:

- *Homogeneity:* $\forall S\in\mathcal{P}(\mathcal{X})\setminus\{\emptyset\},\forall\mathbf{x}\in\mathcal{X},\ \mathop{\mathbb{E}}_{\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}[\psi_S(\mathbf{y})]=\rho^{|S|}\psi_S(\mathbf{x})$

- *Stochastic anisotropy:* $\forall S \in \mathcal{P}(\mathcal{X}) \setminus \{\emptyset\}, \forall \mathbf{x} \in \mathcal{X}, \underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}[\psi_S(\mathbf{y})] = \frac{\rho^{|S|}}{n^l}\psi_S(\mathbf{x})$

PROOF. We define the following set of functions:

$$\forall \mathcal{D}, \forall S \in \mathcal{P}(\mathcal{X}), \forall \mathbf{x} \sim \mathcal{D}, \chi_S(\mathbf{x}) \triangleq \prod_{i\in S}(x_i - \underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i])$$

□

Step 1: Homogeneity and stochastic entropy of the set $\{\chi_S\}_{S\in\mathcal{P}(\mathcal{X})}$.

LEMMA 1. *Homogeneity*

$$\underset{\mathbf{y}\sim\Gamma_{\rho}(\mathbf{x})}{\mathbb{E}}[\chi_S(\mathbf{y})] = \rho^{|S|}\chi_S(\mathbf{x})$$

Let $S \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
\underset{\mathbf{y}\sim\Gamma_{\rho}(\mathbf{x})}{\mathbb{E}}[\chi_S(\mathbf{y})] &= \underset{\mathbf{y}\sim\Gamma_{\rho}(\mathbf{x})}{\mathbb{E}}\left[\prod_{i\in S}\Big(y_i - \underset{\substack{\mathbf{y}\sim\Gamma_{\rho}\\ \mathbf{x}\sim\mathcal{D}}}{\mathbb{E}}[y_i]\Big)\right] \\
&= \underset{\mathbf{y}\sim\Gamma_{\rho}(\mathbf{x})}{\mathbb{E}}\left[\prod_{i\in S}\Big(y_i - \rho\underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i]\Big)\right] \\
&= \prod_{i\in S}\Big(\underset{\mathbf{y}\sim\Gamma_{\rho}(\mathbf{x})}{\mathbb{E}}[y_i] - \rho\underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i]\Big) \\
&= \rho^{|S|}\prod_{i\in S}\Big(x_i - \underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i]\Big) \\
&= \rho^{|S|}\chi_S(\mathbf{x})
\end{aligned}$$

LEMMA 2. *Stochastic anisotropy*

$$\underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}[\chi_S(\mathbf{y})] = \frac{\rho^{|S|}}{n^l}\chi_S(\mathbf{x})$$

Let $S \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
\underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}[\chi_S(\mathbf{y})] &= \underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}\left[\prod_{i\in S}\Big(y_i - \underset{\substack{\mathbf{y}\sim\Gamma_{\rho,l}\\ \mathbf{x}\sim\mathcal{D}}}{\mathbb{E}}[y_i]\Big)\right] \\
&= \underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}\left[\prod_{i\in S}\Big(y_i - \frac{1}{n^l}\rho\underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i]\Big)\right] \\
&= \prod_{i\in S}\Big(\underset{\mathbf{y}\sim\Gamma_{\rho,l}(\mathbf{x})}{\mathbb{E}}[y_i] - \frac{1}{n^l}\rho\underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_i]\Big)
\end{aligned}$$

$$= \frac{\rho^{|S|}}{n^l} \prod_{i \in S} \Big(x_i - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}}[x_i] \Big)$$
$$= \frac{\rho^{|S|}}{n^l} \chi_S(\mathbf{x})$$

Step 2: Basis construction via Gram-Schmidt orthogonalization.
We build the Gram-Schmidt basis from the set of functions $\{\chi_S\}$, using this ordered set:

$$\varnothing, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \ldots, \{1,2,3,\ldots,n\}$$

Let's denote $\{S_i\}_{i \in [\![1, 2^n]\!]}$ those sets ordered bellow, the basis construction follows:

$$\psi_\emptyset = 1$$
$$\forall i \in [\![1, 2^n]\!] : \psi_{S_i} = \chi_{S_i} - \sum_{j=1}^{i-1} \frac{\rho^{|S_i|}}{n^l \rho^{|S_j|}} \langle\, \psi_{S_j}, \chi_{S_i} \rangle_{\mathcal{D}} \psi_{S_j}$$
$$\psi_{S_i} = \begin{cases} \frac{\tilde{\psi}_{S_i}}{\|\tilde{\psi}_{S_i}\|_{2,\mathcal{D}}} & \text{if } \tilde{\psi}_{S_i} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

For all $i \in [\![1, 2^n]\!]$, let $\mathcal{Q}(i)$ denote the following property:
$\mathcal{Q}(i) \triangleq$ " ψ_{S_i} satisfies both homogeneity and stochastic anistropy"
The case where $i = 0$ corresponds to $S = \emptyset$, which is excluded.
Let's prove $\mathcal{Q}(i)$ by strong induction on i. Fix $\mathbf{x} \in \mathcal{X}$, for all $\mathbf{y} \sim \Gamma_\rho$, we have:

$$\psi_{S_1}(\mathbf{y}) = \chi_{S_1}(\mathbf{y})$$

By transition of homogeneity and stochastic anisotropy for the set of functions $\{\chi_S\}$, $\mathcal{Q}(1)$ is true.

We suppose that $\mathcal{Q}(1), \ldots \mathcal{Q}(i)$ are true, let's show $\mathcal{Q}(i+1)$:

Homogeneity: This corresponds to the basis with a single parameter ($l = 0$). From Lemma 1 and the induction assumption, it follows:

$$\mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_\rho(\mathbf{x})}[\psi_{S_{i+1}}(\mathbf{y})] = \mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_\rho(\mathbf{x})}[\chi_{S_{i+1}}(\mathbf{y})] - \sum_{j=1}^{i} \frac{\rho^{|S_{i+1}|}}{\rho^{|S_j|}} \langle\, \psi_{S_j}, \chi_{S_{i+1}} \rangle_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_\rho(\mathbf{x})}[\psi_{S_j}(\mathbf{y})]$$
$$= \rho^{|S_{i+1}|} \psi_{S_{i+1}}(\mathbf{x})$$

Stochastic anisotropy: This corresponds to the parametric basis with $((\rho, l) \neq (0, 0))$. From Lemma 2 and the induction assumption, it follows:

$$\mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_{\rho,l}(\mathbf{x})}[\psi_{S_{i+1}}(\mathbf{y})] = \mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_{\rho,l}(\mathbf{x})}[\chi_{S_{i+1}}(\mathbf{y})] - \sum_{j=1}^{i} \langle\, \psi_{S_j}, \chi_{S_{i+1}} \rangle_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{y} \sim \Gamma_{\rho,l}(\mathbf{x})}[\psi_{S_j}(\mathbf{y})]$$
$$= \frac{\rho^{|S_{i+1}|}}{n^l} \psi_{S_{i+1}}(\mathbf{x})$$

This shows that $\mathcal{Q}(i+1)$ is true, which conclude the proof.

APPENDIX D: COMPUTING PROPERTIES IN TERMS OF FOURIER COEFFICIENTS.

In this section, we provide proofs of results given in section 3, we explain the relationship between our properties of interest and influence functions, next, we express the (deterministic and random) influence functions in terms of the model's Fourier coefficients, before deducing the expression of our properties in their Fourier pattern.

D.1 Robustness

Proof of proposition 4: If $\text{Inf}_\rho(h)$ denotes the ρ-Random influence function, we have the following result that relates it to model h Fourier coefficients:

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

Proof. Let $\rho \in [-1,1]$ and $h : \mathcal{X} \to \{-1,1\}$,
We introduce the noise operator $\mathcal{T}_\rho : \mathcal{X} \to \mathbb{R}$ defined as:

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

For all $\mathbf{x} \in \mathcal{X}$, we have:

$$\begin{aligned}
\mathcal{T}_\rho h(\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}[h(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}[\sum_{S\subseteq[n]} \hat{h}(S)\psi_S(\mathbf{y})] \\
&= \sum_{S\subseteq[n]} \hat{h}(S)\psi_S(\mathbf{x})\rho^{|S|} \\
\forall \mathbf{x} \in \mathcal{X} : \mathcal{T}_\rho h(\mathbf{x}) &= \sum_{S\subseteq[n]} \hat{h}(S)\psi_S(\mathbf{x})\rho^{|S|}
\end{aligned}$$

Where the third step derives from the homogeneity of the basis (Proposition 3).
On the other hand:

$$\begin{aligned}
\text{Inf}_\rho(h) &= \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim\Gamma_\rho(\mathbf{x})}}[h(\mathbf{x})h(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \mathop{\mathbb{E}}_{y\sim\Gamma_\rho(\mathbf{x})} h(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\mathcal{T}_\rho h(\mathbf{x})] \\
&= \langle\, h, \mathcal{T}_\rho h\rangle \\
&= \langle \sum_{S_1\subseteq[n]} \hat{h}(S_1)\psi_{S_1}(.), \sum_{S_2\subseteq[n]} \rho^{|S_2|}\hat{h}(S_2)\psi_{S_2}(.)\rangle \\
&= \sum_{S_1\subseteq[n]}\sum_{S_2\subseteq[n]} \rho^{|S_2|}\hat{h}(S_1)\hat{h}(S_2)\langle\, \psi_{S_1}(.), \psi_{S_2}(.)\rangle
\end{aligned}$$

$$= \sum_{S_1 \subseteq [n]} \sum_{S_2 \subseteq [n]} \rho^{|S_2|} \hat{h}(S_1) \hat{h}(S_2) \delta_{S_1, S_2}$$

Hence,

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

□

We deduce the Fourier pattern in robustness property:

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

D.2 Individual fairness

Proof of proposition 5: The individual fairness discrepancy is given by:

$$\mu_{\mathrm{IFair}}(h) = \sum_{S \subseteq [n]} \frac{\rho^{|S|}}{n^l} \hat{f}(S)^2$$

Proof. First, we introduce the fair operator.
Let $\rho \in [-1, 1]$ and $l \in \mathbb{N}$ such that $l \leq n$ and $h : \{-1, 1\}^n \to \{-1, 1\}$,
We introduce the fair operator $\mathcal{T}_{\rho,l} : \{-1, 1\}^n \to \mathbb{R}$ defined as:

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

Recall the protocol that generates the perturbed vector $\mathbf{y}$ from $\mathbf{x}$, independently:

- The first step, we uniformly choose l coefficients of the vector $\mathbf{x}$ to be perturbed.
- The second step, we generate the vector $\mathbf{y}$, by maintaining the $n-l$ coefficients of the vector $\mathbf{x}$ and concatenate with the perturbed directions. The uniformly random perturbation is defined as (independently):

$$y_i = \begin{cases} x_i & \text{with probability } \frac{1+\rho}{2} \\ -x_i & \text{with probability } \frac{1-\rho}{2} \end{cases}$$

For all $\mathbf{x} \in \{-1, 1\}^n$, we have:

$$\begin{aligned} \mathcal{T}_{\rho,l} h(\mathbf{x}) &= \mathop{\mathbb{E}}_{\mathbf{y} \sim N_{\rho,l}(\mathbf{x})} [h(\mathbf{y})] \\ &= \mathop{\mathbb{E}}_{y \sim N_{\rho,l}(\mathbf{x})} [\sum_{S \subseteq [n]} \hat{h}(S) \psi_S(\mathbf{y})] \\ &= \sum_{S \subseteq [n]} \hat{h}(S) \mathop{\mathbb{E}}_{y \sim N_{\rho,l}(\mathbf{x})} [\psi_S(\mathbf{y})] \end{aligned}$$

$$= \sum_{S\subseteq[n]} \frac{\rho^{|S|}}{n^l}\hat{h}(S)\psi_S(\mathbf{x})$$

$$\forall x \in \{-1,1\} : \mathcal{T}_{\rho,l}h(\mathbf{x}) = \sum_{S\subseteq[n]} \frac{\rho^{|S|}}{n^l}\hat{h}(S)\psi_S(\mathbf{x})$$

Where the fourth step comes from the stochastic anisotropy of the basis (Proposition 3). On the other hand:

$$\begin{aligned}
\mu_{\text{IFair}}(h) &= \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}}[h(\mathbf{x})h(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \mathop{\mathbb{E}}_{\mathbf{y}\sim N_{\rho,l}(\mathbf{x})} h(\mathbf{y})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x})\mathcal{T}_{\rho,l}h(\mathbf{x})] \\
&= \langle\ h, \mathcal{T}_{\rho,l}h\rangle \\
&= \langle \sum_{S_1\subseteq[n]} \hat{h}(S_1)\psi_{S_1}(.), \sum_{S_2\subseteq[n]} \frac{\rho^{|S_2|}}{n^l}\hat{h}(S_2)\psi_{S_2}(.)\rangle \\
&= \sum_{S_1\subseteq[n]}\sum_{S_2\subseteq[n]} \frac{\rho^{|S_2|}}{n^l}\hat{h}(S_1)\hat{h}(S_2)\langle\ \psi_{S_1}(.), \psi_{S_2}(.)\rangle \\
\mu_{\text{IFair}}(h) &= \sum_{S_1\subseteq[n]}\sum_{S_2\subseteq[n]} \frac{\rho^{|S_2|}}{n^l}\hat{h}(S_1)\hat{h}(S_2)\delta_{S_1,S_2}
\end{aligned}$$

Hence,

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

□

D.3 Group Fairness.

We show the relationship between group fairness and Fourier coefficients via the following lemma:

LEMMA 3. *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:

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

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:

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

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

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

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

□

Proof of proposition 6: We use the following notation in the proof:

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

We have,

$$\mu_{\text{GFair}}(h) = \mu_{\text{GFair}}(h)^+ - \mu_{\text{GFair}}(h)^- \tag{D.1}$$

By the law of total probability, we also have:

$$p = \alpha\mu_{\text{GFair}}(h)^+ + (1-\alpha)\mu_{\text{GFair}}(h)^- \tag{D.2}$$

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

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

Hence, we have:

$$\mu_{\text{GFair}}(h)^+ + \mu_{\text{GFair}}(h)^- - 2\mu_{\text{GFair}}(h)^+\mu_{\text{GFair}}(h)^- - \text{Inf}_A(h) = 0$$

From equation D.1, and equation D.2, we have:

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

The expression becomes:

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

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.
Which concludes the proof.

Proof of corollary 2.

If $\mathcal{D}$ is the uniform distribution, SP is exactly the Fourier coefficient of this sensitive attribute:

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

PROOF.

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

□

APPENDIX E: THEORETICAL ANALYSIS.

E.1 NP-hardness of exact computation.

In this section, we prove Theorem 3.

If $\mathcal{Q}$ denotes the set of queries, $\mathcal{H}_{ist} = \{x, f(x)\}_{x\in\mathcal{Q}}$ denotes the history generated from the interaction between a learning algorithm and a black box oracle of our model h via membership queries, and $\tau \in \mathbb{R}$ denotes the threshold, then finding significant Fourier coefficients of h for the threshold τ is NP-complete.

PROOF. The problem of finding large Fourier coefficients of a Boolean function for some threshold can be converted to finding a variable assignment that maximizes the number of clauses of a CNF of two literals with at least one true literal, that is each coordinate of a sample from $\mathcal{X}$ form a clause with exactly two literals (per clause). Since the Fourier coefficient of h is $\hat{h}(S) = \frac{1}{|\mathcal{Q}|}\sum_{x\in\mathcal{Q}} h(\mathbf{x})\psi_S(\mathbf{x})$, one can observe that $h(\mathbf{x})$ in the expression above give the number of clauses that will be satisfied, Hence, finding large Fourier coefficients can be reduced to the problem of finding a variable assignment that maximizes the number of clauses with at least one true literal, which is exactly the Max2Sat problem. □

E.2 Manipulation-proof

Here we prove Proposition 4: AFA achieves optimal manipulation-proofness for estimating statistical parity with manipulation-proof subclass of size 2^{n-2}.

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

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

$$h = \sum_{S\subseteq[n]} \hat{h}(S)\psi_S$$

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

On the other hand,

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

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 6).

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}. □

E.3 Upper bounds.

Robustness and individual fairness (Theorem 5).
Let $\mu \in \{\mu_{\text{RB}}, \mu_{\text{IF}}\}$.
If h denotes the true model, then there exists a PAC-agnostic auditor for μ with query complexity $\mathcal{O}\Big(\frac{8\sqrt{2}\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\Big)$.
Where $\texttt{char}(L,\mu) = \sum_{S\in L}\texttt{char}(S,\mu)$, $\texttt{char}(\bar{L},\mu) = \sum_{S\in\bar{L}}\texttt{char}(S,\mu)$, and L is the list of subsets with significant Fourier coefficients.

Proof. Let $\epsilon, \delta \in (0,1)$, and $\tau^2 \triangleq 4\epsilon$. Let m denote the sample complexity for auditing robustness and individual fairness. we $x_1, \cdots, x_{m_1}, x'_1, \cdots, x'_{m_2}$ m points ($m = m_1 + m_2$) from $\mathcal{X}$ sampled iid from $\mathcal{D}$.

Let L denote the list of subsets satisfying large Fourier coefficients, obtained at the last level of the tree (Figure 3). We consider the following unbiased estimator of squared Fourier coefficients:

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

Hence, the property (robustness and individual fairness) estimator takes the following form:

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

Where the ground truth is given by :

$$\mu(h) = \sum_{S\subseteq[\![1,n]\!]} \texttt{char}(S,\mu)\hat{h}_{\mathsf{AFA}}(S)^2$$

$$= \sum_{S\subseteq[\![1,n]\!]} \texttt{char}(S,\mu)\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}}[h(\mathbf{x})h(\mathbf{y})\psi_S(\mathbf{x})\psi_S(\mathbf{y})]$$

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

Where the last step comes from the correctness of Goldreich-Levin, that is:

$$\forall S\subseteq[\![1,n]\!] : S\notin L \implies |\hat{h}(S)|\leq\tau$$

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 $\sum_{S\notin L}\texttt{char}(S,\mu)$. We deduce:

$$\mathbb{P}\Big[|\mu(h)-\hat{\mu}_{\mathsf{AFA}}|\leq\epsilon\Big]\geq\mathbb{P}\Big[\sum_{S\in L}\texttt{char}(S,\mu)|\hat{h}(S)^2-\hat{h}_{\mathsf{AFA}}(S)^2|\leq\epsilon-\tau^2\texttt{char}(\bar{L},\mu)\Big] \tag{E.1}$$

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

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

The proof is by induction on $|\mathcal{I}|$.

Lemma 4. *Two-Sample Hoeffding If $X_1,...X_{m_1}$, $X'_1,...X'_{m_2}$ are iid random variables taking values in $[-1,1]$ generating by the distribution $\mathcal{D}$, and:*

$$\mu=\mathbb{E}[X^2]$$

,

$$\hat{\mu}=\frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}X_iX'_j$$

then:

$$\mathbb{P}[|\hat{\mu}-\mu|\leq 4\epsilon]\geq 1-2\exp\Big\{-\frac{m_1m_2\epsilon^2}{8}\Big\}$$

The proof is by employing One sample Hoeffding inequality to the random variable $Z_{i,j} = X_i X_j'$. Applying Claim 1,

$$\begin{aligned}\mathbb{P}\Big[\bigcap_{S\in L}\Big\{\Big|\hat{h}(S)^2 - \hat{h}_{\mathsf{AFA}}(S)^2\Big| \leq 4\epsilon\Big\}\Big] &\geq \sum_{S\in L}\mathbb{P}\Big[\Big|\hat{h}(S)^2 - \hat{h}_{\mathsf{AFA}}(S)^2\Big| \leq 4\epsilon\Big] - |L| + 1\\ &\geq |L| - 2|L|\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{8}\Big\} - |L| + 1\\ &\geq 1 - 2|L|\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{8}\Big\}\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{P}\Big[\sum_{S\in L}\mathtt{char}(S,\mu)\Big|\hat{h}(S)^2 - \hat{h}_{\mathsf{AFA}}(S)^2\Big| \leq 4\mathtt{char}(L,\mu)\epsilon\Big] &\geq \mathbb{P}\Big[\bigcap_{S\in L}\Big\{\Big|\hat{h}(S)^2 - \hat{h}_{\mathsf{AFA}}(S)^2\Big| \leq 4\epsilon\Big\}\Big]\\ &\geq 1 - 2|L|\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{8}\Big\}\end{aligned}$$

By a change of variables, and the fact that $\epsilon = \frac{\tau^2}{4}$

$$\mathbb{P}\Big[\sum_{S\in L}\mathtt{char}(S,\mu)\Big|\hat{h}(S)^2 - \hat{h}_{\mathsf{AFA}}(S)^2\Big| \leq \epsilon - \tau^2\mathtt{char}(\bar{L},\mu)\Big] \geq 1 - 2|L|\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{128\mathtt{char}(L,\mu)^2(1-4\mathtt{char}(\bar{L},\mu))^2}\Big\}$$

From inequality E.1,

$$\mathbb{P}\Big[|\mu(h) - \hat{\mu}_{\mathsf{AFA}}| \leq \epsilon\Big] \geq 1 - 2|L|\exp\Big\{-\frac{m_1 m_2 \epsilon^2}{128\mathtt{char}(L,\mu)^2(1-4\mathtt{char}(\bar{L},\mu))^2}\Big\}$$

By the definition of the sample complexity, this gives:

$$\begin{aligned}m_1 m_2 &\geq \frac{128\mathtt{char}(L,\mu)^2(1-4\mathtt{char}(\bar{L},\mu))^2}{\epsilon^1}\log\frac{2|L|}{\delta}\\ &\geq \frac{128\mathtt{char}(L,\mu)^2(1-4\mathtt{char}(\bar{L},\mu))^2}{\epsilon^2}\log\frac{2}{\delta}\end{aligned}$$

Since $m = m_1 + m_2 \geq 2\sqrt{m_1 m_2}$, we get:

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

□

Group fairness.
From Proposition 6, we have the discriminant of the second order polynomial takes the form:

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

For $\epsilon > \frac{1}{8}$, Δ is positive.
Note: This condition is independent of α or p, unlike other assumptions in previous work [YZ22].
μ_{Rob}, the zero of a second-order polynomial, which can be expressed as:

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

We consider the following estimator:

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

Where,

$$
\begin{aligned}
p = p(h) = \frac{1+\hat{h}(\emptyset)}{2}, \quad \hat{p} = \frac{1+\hat{h}_{\mathsf{AFA}}(\emptyset)}{2} \\
\mathrm{Inf}_A = \mathrm{Inf}_A(h) = \sum_{\substack{S\subseteq[\![1,n]\!] \\ S\ni A}} \hat{h}(S)^2, \quad \widehat{\mathrm{Inf}_A} = \sum_{\substack{S\in L \\ S\ni A}} \hat{h}_{\mathsf{AFA}}(S)^2
\end{aligned}
$$

To simplify notations, we denote:

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

$$
\hat{\Delta} = 4\hat{\alpha}^2 + 4\hat{p}^2 - 4\alpha - 4p + 8\alpha(1-\alpha)\widehat{\mathrm{Inf}_A} + 1 \tag{E.3}
$$

We have,

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

On the other hand,

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

Similar to the previous proof and by using Two-sample Hoeffding on the first and third term above, we get a sample complexity upper bound of $\mathcal{O}\left(\max\left\{\frac{1}{\epsilon^2}\log\frac{4}{\delta}, \frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}\right\}\right)$.

E.4 Lower bounds.

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

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} - \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} + \epsilon \end{cases} \tag{E.4}$$

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

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} + \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} - \epsilon \end{cases} \tag{E.5}$$

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

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} + \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} - \epsilon \end{cases} \tag{E.6}$$

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

$$h^*(z_i) := \begin{cases} 1 & \text{with probability} \quad \frac{1}{2} - \epsilon \\ 0 & \text{with probability} \quad \frac{1}{2} + \epsilon \end{cases} \tag{E.7}$$

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

The case where $\text{VC}(\mathcal{H}) \in 2\mathbb{N}+1$ reduces to $\text{VC}(\mathcal{H}) \in 2\mathbb{N}$ 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.

$$\mathbb{P}\left[\hat{\mu} - \mu(h^*) > \epsilon\right] \geq \underbrace{\mathbb{P}\left[\hat{p} - p(h^*) > c_\alpha \epsilon\right]}_{\text{Term I}} + \underbrace{\mathbb{P}\left[\hat{\mathrm{Inf}}_A - \mathrm{Inf}_A(h^*) > c_{\alpha,1}\epsilon^2 + c_{\alpha,2}\right]}_{\text{Term II}}, \tag{E.8}$$

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:

$$\begin{aligned}
\mathbb{P}\left[\hat{p} - p(h^*) \geq \frac{\epsilon}{2}\right] \geq \frac{1}{2}\Bigg(&\mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha-1}{2}\right] + \mathbb{P}_1\left[\hat{p} < \frac{2\alpha-1}{2}\right] + && \text{(E.9)} \\
&\mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\right] + \mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha-1}{2} + \frac{\epsilon}{2}\right] - 2\Bigg) && \text{(E.10)}
\end{aligned}$$

By Le Cam's lemma:

$$\mathbb{P}_0\left[\hat{p} \geq \frac{2\alpha-1}{2}\right] + \mathbb{P}_1\left[\hat{p} < \frac{2\alpha-1}{2}\right] \geq 1 - \mathrm{TV}\left(\mathbb{P}_0 \parallel \mathbb{P}_1\right) \tag{E.11}$$

Concentration of $p(h^*)$. To lower bound the remaining term in E.10, we prove the following lemma:

LEMMA 5.

$$\mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{E.12}$$

$$\mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha-1}{2} + \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{E.13}$$

PROOF.

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

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. $\frac{1}{2}-\epsilon$).

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

On the other hand,

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

□

This proves result E.14.
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 E.15 concludes by proceeding with the remaining steps in the same manner as the previous proof.

$$\mathbb{P}_0\left[p(h^*) \leq \frac{2\alpha-1}{2} - \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{E.14}$$

$$\mathbb{P}_1\left[p(h^*) \geq \frac{2\alpha-1}{2} + \frac{\epsilon}{2}\right] \geq 1 - 2\exp\Big(-\frac{d\epsilon^2}{32\alpha^2}\Big) - 2\exp\Big(-\frac{d\epsilon^2}{2(1-\alpha)^2}\Big) \tag{E.15}$$

By symmetry of the statistical test we have the result in E.15.

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

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

We deduce

$$\mathbb{P}\left[\hat{\mathrm{Inf}}_A - \mathrm{Inf}_A(h^*) \geq \frac{\epsilon}{2}\right] \geq \frac{1}{2}\Bigg(\mathbb{P}_0\left[\hat{\mathrm{Inf}}_A \geq \frac{1}{2}\right] + \mathbb{P}_1\left[\hat{\mathrm{Inf}}_A < \frac{1}{2}\right] + \tag{E.16}$$

$$\mathbb{P}_0\left[\mathrm{Inf}_A(h^*) \leq \frac{1}{2} - \frac{\epsilon}{2}\right] + \mathbb{P}_1\left[\mathrm{Inf}_A(h^*) \geq \frac{1}{2} + \frac{\epsilon}{2}\right] - 2\Bigg) \tag{E.17}$$

By Le Cam's lemma, we have:

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

Concentration of $\mathrm{Inf}_A(h^*)$. To lower bound the remaining term in E.17, we prove the following lemma:

Under hypothesis H_0, we have:

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

Under hypothesis H_1, we have:

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

Concentration of Influence Function.

LEMMA 6.

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

$$\mathbb{P}_1\Big[\mathrm{Inf}_A(h^*) > \frac{1-\epsilon}{2}\Big] \geq 3 - 4\exp\Big(-\frac{d\epsilon}{2}\Big) - 4\exp\Big(-\frac{d\epsilon}{18}\Big) \tag{E.19}$$

Proof:

$$\begin{aligned}
\mathrm{Inf}_A(h^*) &= \mathbb{P}\left[h^*(x) \neq h^*(x') \middle| x \in \mathcal{X}^+, x' \in \mathcal{X}^-\right] \\
&= \mathbb{P}\left[h^*(x) = 1, h^*(x') = 0 \middle| x \in \mathcal{X}^+, x' \in \mathcal{X}^-\right] + \mathbb{P}\left[h^*(x) = 0, h^*(x') = 1 \middle| x \in \mathcal{X}^+, x' \in \mathcal{X}^-\right] \\
&= \mathbb{P}\left[h^*(x) = 1 \middle| x \in \mathcal{X}^+\right] \mathbb{P}\left[h^*(x) = 0 \middle| x \in \mathcal{X}^-\right] + \mathbb{P}\left[h^*(x) = 0 \middle| x \in \mathcal{X}^+\right] \mathbb{P}\left[h^*(x) = 1 \middle| x \in \mathcal{X}^-\right] \\
&= \frac{1}{d^2} \sum_{1 \leq i,j \leq d} \mathbb{1}_{\{h^*(z_i)=1\}} \mathbb{1}_{\{h^*(z_{d+j})=0\}} + \frac{1}{d^2} \sum_{1 \leq i,j \leq d} \mathbb{1}_{\{h^*(z_i)=0\}} \mathbb{1}_{\{h^*(z_{d+j})=1\}} \\
\mathrm{Inf}_A(h^*) &= \mathrm{Inf}^+_{A,1}(h^*)\mathrm{Inf}^-_{A,0}(h^*) + \mathrm{Inf}^+_{A,0}(h^*)\mathrm{Inf}^-_{A,1}(h^*)
\end{aligned}$$

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:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,1}(h^*) > \frac{1}{2} - \frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{E.20}$$

$$\mathbb{P}_0\left[\mathrm{Inf}^-_{A,0}(h^*) > \frac{1}{2} - \frac{\epsilon}{2}\right] \leq 2\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{E.21}$$

From E.20 and E.21, we deduce:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,1}(h^*)\mathrm{Inf}^-_{A,0}(h^*) \leq \Big(\frac{1}{2} - \frac{\epsilon}{2}\Big)^2\right] \geq 2 - 4\exp\Big(-\frac{d\epsilon^2}{2}\Big) \tag{E.22}$$

Similar, the upper bound of the second part is:

$$\mathbb{P}_0\left[\mathrm{Inf}^+_{A,0}(h^*)\mathrm{Inf}^-_{A,1}(h^*) \leq \left(\frac{1}{2}+\frac{\epsilon}{2}\right)^2\right] \geq 2-4\exp\left(-\frac{d\epsilon^2}{18}\right) \tag{E.23}$$

Combining results E.22 andE.23 yields result E.18. By the symmetry of the hypotheses H_0 and H_1, we obtain the second result.

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 have the following lemma:

Lemma 7.

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

Proof. By definition,

$$\begin{aligned}
\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1) &= \sum_{\mathcal{Q}\in\mathcal{H}^{ist}_N}\mathbb{P}_0(\mathcal{Q})\log\frac{\mathbb{P}_0(\mathcal{Q})}{\mathbb{P}_1(\mathcal{Q})}\\
&= \sum_{\substack{\mathcal{Q}\in\mathcal{H}^{ist}_N\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_N,y_N)\}}}\mathbb{P}_0(\mathcal{Q})\log\frac{\prod_{i=1}^N\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\mathbb{P}_{\mathcal{A}}(x_i|(x,y)\in\mathcal{H}^{ist}_{i-1})}{\prod_{i=1}^N\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\mathbb{P}_{\mathcal{A}}(x_i|(x,y)\in\mathcal{H}^{ist}_{i-1})}\\
&= \sum_{\substack{\mathcal{Q}\in\mathcal{H}^{ist}_N\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_N,y_N)\}}}\mathbb{P}_0(\mathcal{Q})\sum_{i=1}^N\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}\\
&= \sum_{i=1}^N\sum_{\substack{\mathcal{Q}\in\mathcal{H}^{ist}_N\\ \mathcal{Q}=\{(x_1,y_1),\dots(x_i,y_i)\}}}\mathbb{P}_0(\mathcal{Q})\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}\\
&= \sum_{i=1}^N\sum_{\{(x_1,y_1),\dots(x_i,y_i)\}}\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\mathbb{P}_0((x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}\\
&= \sum_{i=1}^N\sum_{\{(x_1,y_1),\dots(x_{i-1},y_{i-1}),x_i\}}\mathbb{P}_0((x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\sum_{y_i}\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)\log\frac{\mathbb{P}_0(y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}{\mathbb{P}_1((y_i|(x,y)\in\mathcal{H}^{ist}_{i-1},x_i)}\\
&= \sum_{i=1}^N\sum_{\mathbf{H}_{i-1},x_i}\mathbb{P}_0((x,y)\in\mathbf{H}_{i-1},x_i)\mathcal{D}_{KL}\left(\mathbb{P}_0(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\middle|\middle|\mathbb{P}_1(y_i|(x,y)\in\mathbf{H}_{i-1},x_i)\right)
\end{aligned}$$

Hence,

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

□

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:

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

Claim 2. *For $a,b\in(\frac{1}{4},\frac{3}{4})$: $kl(a,b)\leq 3(b-a)^2$*

Hence,

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

$$\mathcal{D}_{KL}(\mathbb{P}_0||\mathbb{P}_1)\leq 3N(1-2\alpha)^2\epsilon^2 \tag{E.24}$$

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 E.24,

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

Results E.14 and E.15 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 $M_\alpha = \max(\alpha, 1-\alpha)$.
Results E.18 and E.19 yield:

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

solving the inequality:

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

gives $N \leq \frac{8}{3(1-2\alpha)^2\epsilon^2}\left(\delta - 3 + 4\exp\left(-\frac{d\epsilon^2}{18}\right)\right)^2$. □

APPENDIX F: EXTENSION TO CATEGORICAL DOMAIN

We are interested in a more general case, where the input space can have any categorical structure extending to the torus $\mathbb{F}_p^n$. In other words, we're interested in the set of classifiers defined on categorical domain $\mathbb{F}_p^n$, where $p \in \mathbb{N}$ is any number $p > 2$ (not necessarily prime).

Let $p \in \mathbb{N}$ such that $p > 2$, and $\mathbb{F}_p$ denotes the cyclic group $\mathbb{Z}\Big/_{p\mathbb{Z}}$ of order p. We assume now that $\mathcal{X} = \mathbb{F}_p^n$. Similarly to the case of the Boolean domain, function h admits a unique Fourier representation as follows:

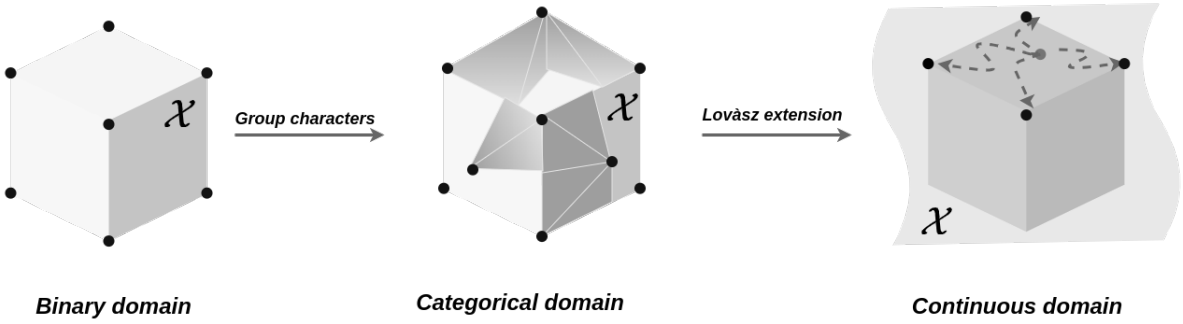

FIG 5. *Geometric intuition: This figure summarizes our approach: First, we study the problem in the case of the binary domain, then we extend the study to obtain the influence for the categorical domain.*

$$h(\mathbf{x}) = \sum_{\zeta} \hat{h}(\zeta)\omega_p(\langle \zeta, x\rangle) \tag{F.1}$$

In this case, the basis becomes the set of unity roots functions $\omega_p(\langle \zeta, \cdot\rangle)_{\zeta\in\mathbb{F}_p^n}$. This set is an orthonormal basis of the vector space $\mathcal{F}_p^n$. The Fourier coefficients are given by:

$$\hat{h}(\zeta) = \mathbb{E}[h(\mathbf{x})\omega_p((\langle \zeta, x\rangle))]$$

In the previous sections, we analyzed the discrepancy fairness property of the model classes embedded in the functional set $\mathcal{F}_2^n = \left\{ f : \mathbb{F}_2^n \to \{0,1\}; \text{f is bounded} \right\}$, we're now interested in a more general case, where the input space can have any categorical structure extending to the torus $\mathbb{F}_p^n$. In other words, we're interested in the functional set $\mathcal{F}_p^n = \left\{ f : \mathbb{F}_p^n \to \{0,1\}; \text{f is bounded} \right\}$, where $p \in \mathbb{N}$ is any number $p > 2$ (not necessarily prime).

We consider the case where a sensitive attribute takes multiple values and we want to estimate the fairness discrepancy of a binary classifier $f \in \mathcal{F}_p^n$, by embedding this unfairness gap in a circle centered in the influence parameter given by Fourier coefficients derived from the model.

An extension to any categorical feature space

we now analyze the discrepancy fairness property of the model classes embedded in the functional set $\mathcal{F}^n_{\{p_1,\dots,p_n\}} = \left\{ h : \prod_{i=1}^n \mathbb{F}_{p_i}^n \to \{0,1\}; \right\}$, In this case, the input space can have categories with different cardinal extending to the product torus $\prod_{i=1}^n \mathbb{F}_{p_i}^n$. First, we define a random flipping operator σ_j as the following action on x: $\sigma_j(x) = (x_1, \dots, x_j + \mathcal{T}_j, \dots, x_n)$.

F.0.1 Robustness Following the same notations in the section, we extend the definitions for the Boolean domain over categorical domains. Let $\rho \in [0,1]^n$, such that $||\rho|| = 1$. For a fixed $x \in \{-1,1\}^n$, we write $y \sim N_p(x)$ to denote the random variable such that for all $i \in [n]$, independently, with probability ρ_i, $y_i = x_i + \mathcal{T}_i$ Where $\mathcal{T}_i$ is the random flipping operator. We'll say that the random variables X and Y are a ρ-correlated pair if $Y \sim N_\rho(X)$.

DEFINITION 10 (ρ-flipping influence function). The ρ-flipping influence function of any model h is defined as:

$$\mathrm{Inf}_\rho(h) = \mathop{\mathbb{E}}_{(x,y)\sim\rho-correlated}[h(x)h(y)]$$

The goal is to estimate the robustness of our model h, defined in a categorical domain, using its Fourier coefficients.

THEOREM 8 (Robustness for categorical domain). *Robustness of h under the N_ρ flipping perturbation is equivalent to the ρ-flipping influence function,*

$$\mu_{\text{Rob}}(h) = (\frac{1}{1-\rho})^n \sum_{\zeta:\zeta_j\neq 0} |\hat{h}(\zeta)|^2 + \sum_{\zeta:\zeta_j=0} |\hat{h}(\zeta)|^2$$

F.0.2 Individual fairness For individual fairness, and particularly for this section of the categorical domain, we consider another measure:

DEFINITION 11 ((ρ,l)-flipping influence function). (ρ,l)-flipping influence function is defined as $\text{Inf}_{\rho,l}(h) = \mathop{\mathbb{E}}_{y\sim N_{\rho,l}(x)}[h(x)h(y)]$

THEOREM 9 (Individual fairness for categorical domain). *The formulation of individual unfairness within the categorical domain is presented by the following expression:*

$$\mu_{\text{IFair}}(h) = \frac{1}{p}\sum_{\zeta} |\hat{h}(\zeta)|^2 cos\Big(\frac{2\pi}{p(1+\rho)l} \sum_{\mathcal{T}\in\mathbb{F}_p^n} \langle\, \zeta, \mathcal{T}\rangle\Big)$$

F.0.3 Group fairness

DEFINITION 12 (ρ-random flipping influence function). Let $h \in \mathcal{F}_p^n$, let $j \in \{1,\dots,n\}$ correspond to the coordinate of the protected attribute, and let $\mathcal{T}_j$ the random variable taking values in $\mathbb{F}_p \setminus \{0\}$. ρ-random flipping influence function is: $\text{Inf}_j(h) = \mathbb{P}[f(X) \neq h(\sigma_j(x))]$

PROPOSITION 7. *Let $h \in \mathcal{F}_p^n$,if j denotes the index of the protected attribute, statistical parity can be expressed as: $\mu_{\text{GFair}}(h) = \frac{2p}{p-1} \sum_{\zeta:\zeta_j\neq 0} |\hat{h}(\zeta)|^2$, which is equivalent to: $\mu_{\text{GFair}}(h) = \frac{2p}{p-1}\sum_{\zeta} |supp(\zeta)||\hat{h}(\zeta)|^2$.*

Proofs for Categorical domain

PROOF. Let $f \in \mathcal{F}_p^n$,
Since h takes values 0 and 1, the influence can be expressed as follows:

$$I_j(f) = \mathbb{E}\left[\left(f(X) - f(\sigma_j(X))\right)^2\right]$$

By Parseval identity:

$$\mathbb{E}[f(X)^2] = \mathbb{E}[f(\sigma_j(X))^2] = \sum_{\zeta} |\hat{f}(\zeta)|^2$$

$$\begin{aligned} I_j(f) &= \mathbb{E}\left[\left(f(X) - f(\sigma_j(X))\right)^2\right] \\ &= \mathbb{E}[f(X)^2] + \mathbb{E}[f(\sigma_j(X))^2] - 2\mathbb{E}[f(X)f(\sigma_j(X))] \qquad \text{(F.2)} \\ I_j(f) &= 2\sum_{\zeta} |\hat{f}(\zeta)|^2 - 2\mathbb{E}[f(X)f(\sigma_j(X))] \end{aligned}$$

$$\begin{aligned}
\mathbb{E}[f(X)f(\sigma_j(X))] &= \mathbb{E}\left[\sum_{\zeta}\sum_{\eta}\hat{f}(\zeta)\overline{\hat{f}}(\eta)\omega_p(\langle\zeta,x\rangle)\omega_p(-\langle\eta,\sigma_j(x)\rangle)\right] \\
&= \sum_{\zeta}\sum_{\eta}\hat{f}(\zeta)\overline{\hat{f}}(\eta)\mathbb{E}\left[\omega_p(\langle\zeta-\eta,x\rangle)\omega_p(-\mathcal{T}_j\eta_j)\right] \\
&= \sum_{\zeta}\sum_{\eta}\hat{f}(\zeta)\overline{\hat{f}}(\eta)\mathbb{E}\left[\delta_{\zeta,\eta}\omega_p(-\mathcal{T}_j\eta_j)\right] \\
\mathbb{E}[f(X)f(\sigma_j(X))] &= \sum_{\zeta}|\hat{f}(\zeta)|^2\mathbb{E}\left[\omega_p(-\mathcal{T}_j\zeta_j)\right]
\end{aligned} \tag{F.3}$$

The expectation in the right part can be expressed as

$$\mathbb{E}\left[\omega_p(-\mathcal{T}_j\zeta_j)\right] = \begin{cases} 1 & \text{if } \zeta_j = 0 \\ \frac{1}{1-p} & \text{otherwise} \end{cases} \tag{F.4}$$

Plugging this in equation F.3:

$$\mathbb{E}[f(X)f(\sigma_j(X))] = \sum_{\zeta:\zeta_j=0}|\hat{f}(\zeta)|^2 + \frac{1}{1-p}\sum_{\zeta:\zeta_j\neq 0}|\hat{f}(\zeta)|^2 \tag{F.5}$$

Equation F.6 gives

$$\begin{aligned}
I_j(f) &= 2\sum_{\zeta}|\hat{f}(\zeta)|^2 - 2\sum_{\zeta:\zeta_j=0}|\hat{f}(\zeta)|^2 + 2\frac{1}{p-1}\sum_{\zeta:\zeta_j\neq 0}|\hat{f}(\zeta)|^2 \\
&= 2\sum_{\zeta:\zeta\neq 0}|\hat{f}(\zeta)|^2 + \frac{2}{p-1}\sum_{\zeta:\zeta_j\neq 0}|\hat{f}(\zeta)|^2 \\
I_j(f) &= \frac{2p}{p-1}\sum_{\zeta:\zeta_j\neq 0}|\hat{f}(\zeta)|^2
\end{aligned} \tag{F.6}$$

□

F.1 Computational hardness of influence functions estimation

Similarly to the previous setting of Boolean domain, the Fourier coefficients of our model are given by the following formula:

$$\forall\zeta\in\mathbb{F}_p^n : \hat{f}(\zeta) = \mathbb{E}[f(x)\omega_p((\langle\zeta,x\rangle))]$$

A single Fourier coefficient can be computed in $\mathcal{O}(|\mathcal{X}|)$ time. On the other hand, the cardinal of Fourier coefficients to compute in order to estimate our model's properties is exponential to the size of the input domain; there are p^n Fourier coefficients to compute, hence to compute our model properties it requires finding large Fourier coefficients which are computed in total time $\mathcal{O}(p^n|\mathcal{X}|)$.

Property estimation using influence functions is hard

The following theorem states that learning significant (large with respect to some threshold) Fourier coefficients is NP-hard for the categorical domain.

THEOREM 10. *Let $\mathcal{H}^{ist} = \{x, f(x)\}_{x\in\mathcal{Q}}$ denotes the history generated from the Membership query algorithm which interacts with the black box oracle of our model h. Given a threshold $\tau \in \mathbb{R}$, the decision problem of finding significant Fourier coefficients with respect to the threshold τ is NP-complete.*

In other words, testing the existence of $S \subseteq [n]$ for which $|\hat{f}(S)| > \tau$ is NP-complete.

PROOF. The proof is similar to the Boolean case, except in the categorical case with p categories, finding large coefficients can be reduced to MaxpSAT problem, by considering CNF with p literals. □

Extension to continuous domains: We may further extend our discussion by employing interpolation in the continuous domain, using the Lovász extension of Boolean functions to the simplex $\mathcal{X} = [0, 1]^n$. This approach allows us to derive continuous formulations of our properties of interest. While a detailed analysis of this extension is beyond the scope of this paper, interested readers can refer to the Lovász extension (including concave and convex interpolation in the hypercube). Similarly, the Fourier expansion in the continuous case, though outside the scope of this paper, reflects the Boolean case with minor modifications. Specifically, the Fourier expansion is given by:

$$h(\mathbf{x}) = \sum_{S\subseteq[n]} \hat{h}(S)\psi_S(\mathbf{x})$$

where $\psi_S(\mathbf{x}) = \psi_S(x_1, \ldots, x_n) = \min_{i\in S} x_i$.

APPENDIX G: EXPERIMENTAL DETAILS AND RESULTS

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.

G.1 Uniformly random (I.I.D.) estimators (`Uniform`)

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

The true robustness is defined as:

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

Random estimator samples i.i.d points from $\mathcal{D}$, we'll denote the set of samples S, $S \sim, \mathcal{D}$:

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

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

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

Group Fairness: Let S^+ denotes i.i.d samples from the first protected group and S^- i.i.d samples from the second protected group. Group Fairness (with demographic parity measure) is defined as:

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

G.2 Other baselines

We assess our auditor AFA on statistical parity by comparing its performance in sample complexity and running time to the methodologies investigated by [YZ22]. In their method, auditing entails an additional step: approximating the model through reconstruction prior to plugging in the estimator. Those methodologies use active learning algorithms for approximating the black-box model i.e, CAL algorithm [CAR94], along with its variant for property active estimation μ-CAL, and its randomized counterpart. Furthermore, the inefficient AFA algorithm is employed to find significant Fourier coefficients within subsets containing the protected attribute, this model forces search over within subsets containing the protected attribute, characterized by a combinatorial complexity of 2^{n-1}, where n is the dimension of the input space.

G.3 Scalability Results

Table 3 demonstrates the computation time of different methods in estimating group fairness of different ML models. `Uniform` is the fasted among all methods in estimating group fairness, with the cost of higher estimation error (Table 2) On the other hand, `CAL` and their variants incur higher computational cost and cannot even scale for MLP and RF. Our method AFA requires reasonable computational time to yield a more accurate estimate of statistical parity and can scale across different models. *Therefore,* AFA *appears to be the most practical auditor for group fairness compared to baselines.*

TABLE 3
Computation time for estimating statistical parity by different methods.

Model	AFA	`CAL`	μ`CAL`	Randomized μ`CAL`	Non-efficient AFA	`Uniform`
Logistic Regression	29.7	56.2	45.6	47.4	1671	0.1051
MLP	48.1	—	—	—	1746.4	0.144
Random Forest	27.5	—	—	—	1442.7	0.102

G.4 Experimental validation for continuous domain extension.

To evaluate our model's performance on continuous normalized data, we extended our experiment to the student performance dataset, gender is the sensitive attribute. The following figure demonstrates that our algorithm outperforms the uniform estimator baseline.

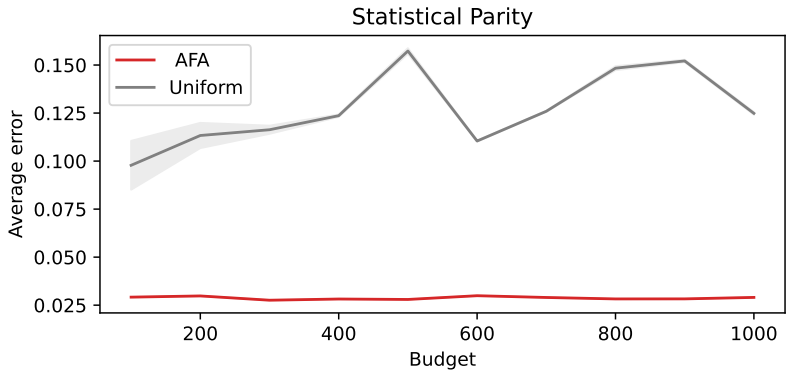

FIG 6. *Error of* AFA *and uniform estimator auditors in estimating statistical parity.*

Figure 7 shows that our algorithm efficiently utilizes memory to compute Fourier weights, enabling the identification of significant Fourier coefficients.

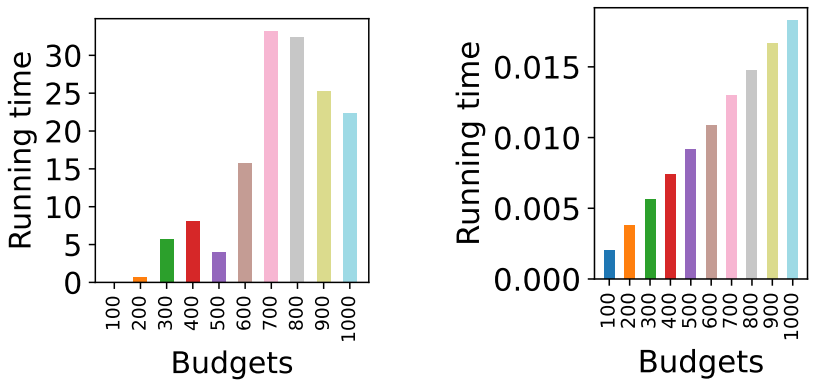

FIG 7. *Running time for* AFA *(left) and for uniform estimator (right) in estimating statistical parity.*

EQUIPE SCOOL
UNIV. LILLE, INRIA
UMR 9189 - CRISTAL, CNRS, CENTRALE LILLE
LILLE, FRANCE

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS
SAARBRUCKEN, GERMANY

EQUIPE SCOOL
UNIV. LILLE, INRIA
UMR 9189 - CRISTAL, CNRS, CENTRALE LILLE
LILLE, FRANCE