Active Fourier Auditor for Estimating Distributional Properties of ML Models

Ayoub Ajarra[1], **Bishwamittra Ghosh**[2], **Debabrota Basu**[1]

[1]Équipe Scool, Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189- CRIStAL, Lille, France
[2]Max Planck Institute for Software Systems, Germany
ayoub.ajarra@inria.fr, bghosh@mpi-sws.org, 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.

Code — https://github.com/ayoubajarra/afamp
Extended version — https://arxiv.org/abs/2410.08111

1 Introduction

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

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

Related Work: Towards trustworthy ML, several methods have been proposed to ally audit an ML model by estimating different *distributional properties* of it, such as fairness and robustness, where the model hyper-property has to be assessed against the distribution of inputs. A stream of work focuses on property verification that verifies whether these properties are violated above a pre-determined threshold (Goldwasser et al. 2021; John, Vijaykeerthy, and Saha 2020; Mutreja and Shafer 2023; Herman and Rothblum 2022; Kearns et al. 2018). Thus, we focus on estimating these properties instead of a 'yes/no' answer, which is a harder problem than verification (Goldwasser et al. 2021). On estimating distributional properties, (Neiswanger, Wang, and Ermon 2021) proposed a Bayesian approach for estimating properties of black-box optimizers and required a prior distribution of models. (Wang et al. 2022) studies simpler distributional properties, e.g. the mean, the median, and the trimmed mean defined as a conditional expectation, using offline and interactive algorithms. (Yan and Zhang 2022) considered a frequentist approach for estimating group fairness but assumed the knowledge of the model class and a finite hypothesis class under audit. These assumptions are violated if we do not know the model type and can be challenging for complex models, e.g. deep neural networks. (Al-

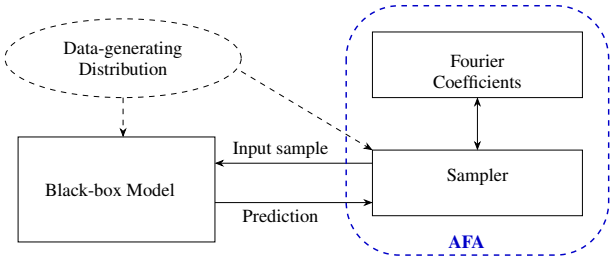

Figure 1: A schematic of AFA.

barghouthi et al. 2017; Ghosh, Basu, and Meel 2021) considered finite models for estimating group fairness w.r.t. the features distribution, and (Ghosh, Basu, and Meel 2022) further narrowed down to linear models. Therefore, we identify the following limitations of the existing methods in ML auditing. (1) **Property-specific auditing:** most methods considered a property-specific tailored approach to audit ML systems, for example either robustness (Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019), group fairness (Albarghouthi et al. 2017; Ghosh, Basu, and Meel 2021), or individual fairness (John, Vijaykeerthy, and Saha 2020). (2) **Model-specific auditing:** all the methods considered a prior knowledge about the ML model (Neiswanger, Wang, and Ermon 2021; Ghosh, Basu, and Meel 2021, 2022; Yan and Zhang 2022), or a white-box access to it (Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019). These are unavailable in practical systems such as API-based ML. Therefore, our research question is: *Can we design a unified ML auditor for black-box systems for estimating a set of distributional properties including robustness and fairness?*

Contributions: We propose a framework, namely AFA (**A**ctive **F**ourier **A**uditor), which is an ML auditor based on the Fourier approximation of a black-box ML model (Figure 1). We observe that existing black-box ML auditors work in two steps: *the model reconstruction step*, where they reconstruct a model completely, and *the estimation step*, where they put an estimator on top of it (Yan and Zhang 2022). We propose a model-agnostic strategy that does not need to reconstruct the model completely. Our contributions are:

- *Formalism.* For any bounded output model (e.g. all classifiers), we theoretically reduce the estimation of robustness, individual fairness, and group fairness in terms of the Fourier coefficients of the model. The key idea is based on influence functions, which capture how much a model output changes due to a change in input variables and can be computed via Fourier coefficients (Section 3). We propose two types of influence functions for each of these properties that unifies robustness and individual fairness auditing while put group fairness in a distinct class.
- *Algorithm.* In AFA, we integrate Goldreich-Levin algorithm (Goldreich and Levin 1989; Kushilevitz and Mansour 1993) to efficiently compute the significant Fourier coefficients of the ML model, which are enough to compute the corresponding properties. AFA yields a probably approximately correct (PAC) estimation of distributional properties. We propose a dynamic version of Goldreich-Levin to accelerate the computations.
- *Theoretical Sample Complexity.* We show that our algorithm requires $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\sqrt{\log\frac{1}{\delta}}\right)$ samples to yield (ϵ,δ) estimate of robustness and individual fairness, while it needs $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ samples to audit group fairness. We further derive a lower bound on the sample complexity of (ϵ,δ)-auditing of group fairness to be $\tilde{\Omega}(\frac{\delta}{\epsilon^2})$. Further, for group fairness, we prove that AFA is manipulation-proof under perturbation of 2^{n-1} Fourier coefficients.
- *Experimental Results.* We numerically test the performance of AFA to estimate the three properties of different types of models. The results show that AFA achieves lower estimation error while estimating robustness and individual fairness across perturbation levels. Compared to existing group fairness auditors, AFA not only achieves lower estimation error but also incurs lower computation time across models and the number of samples.

2 Background

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

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

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

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

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

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

Definition 2.1 (Robustness)**.** Given a model h and a perturbation mechanism Γ of input $x\in\mathcal{X}$, robustness of h is $\mu_{\mathrm{Rob}}(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\ \mathbf{y}\sim\Gamma(x)}[h(\mathbf{x})\neq h(\mathbf{y})]$.

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

In trustworthy and responsible AI, another prevalent concern about deploying ML models is bias in their predic-

tions. This has led to the study of different fairness metrics, their auditing algorithms, and algorithms to enhance fairness (Mehrabi et al. 2021; Barocas, Hardt, and Narayanan 2023). There are two categories of fairness measures (Barocas, Hardt, and Narayanan 2023). The first is the **individual fairness** that aims to ensure that individuals with similar features should obtain similar predictions (Dwork et al. 2012).

Definition 2.2 (Individual Fairness)**.** For a model h and a neighbourhood $\Gamma(x)$ of a $x \in \mathcal{X}$, the individual fairness discrepancy of h is $\mu_{\text{IFair}}(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\ \mathbf{y}\sim\Gamma(\mathbf{x})}\,\mathbb{P}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

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

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

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

Definition 2.3 (Statistical Parity)**.** The statistical parity of h is $\mu_{\text{GFair}}(h) \triangleq |\,\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = 1] - \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = -1]|$, where x_A is the binary sensitive attribute.

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

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

Proposition 2.4 ((Heidari et al. 2021))**.** *There exists a set of orthonormal parity functions $\{\psi_S\}_{S\subseteq[n]}$ such that any function $h : \{-1, 1\}^n \to \{-1, 1\}$ is decomposed as*

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

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

Example 2. Suppose random variables X_1 and X_2 are drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, 1)$ (Heidari et al. 2021). Define another random variable X_3 as $X_3 = X_1X_2$. It can be verified that the Gram-Schmidt basis of XOR of X_1, X_2, X_3 has four zero coefficients, i.e. the sets including X_3 do not influence the outcomes. This is because X_3's information is encoded in X_1 and X_2 jointly.

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

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

Definition 2.5 (Influence functions)**.** If Γ is a transformation of an input $\mathbf{x} \in \mathcal{X}$, the influence function is defined as $\text{Inf}_\Gamma(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \neq h(\Gamma(\mathbf{x}))]$. $\text{Inf}_\Gamma(h)$ is called *deterministic* if the transformation Γ is deterministic, and *randomized* if Γ randomized.

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

3 Active Fourier Auditor

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

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

Definition 3.1 (PAC-agnostic auditor)**.** Let μ be a computable distributional property of model h. An algorithm $\mathcal{A}$ is a *PAC-agnostic auditor* if for any $\epsilon, \delta \in (0, 1)$, there exists a function $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$ samples drawn from $\mathcal{D}$, it outputs an estimate $\hat{\mu}_m$ satisfying $\mathbb{P}(|\hat{\mu}_m - \mu| \leq \epsilon) \geq 1 - \delta$.

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

3.1 The Cost of Reconstruction

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

Proposition 3.2. *If $\hat{h}$ is the reconstructed model from h, then*

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

Proposition 3.2 connects the estimation error and the reconstruction error before plugging in the estimator. It also shows that to have a sensible estimation the reconstruction algorithm needs to achieve an error below the proportion of minority group, which can be significantly small requiring high sample complexity. The proof is deferred to Appendix B. This motivates an approach that avoids model reconstruction by computing only the right components of the model expansion. To capture the information relevant to estimating our properties of interest, we will represent them in terms of Fourier coefficients given in the model decomposition.

3.2 Model Properties with Fourier Expansion

Throughout the rest of this paper, we denote by $\{\psi_S\}_{S\subseteq[n]}$ the basis derived from Proposition 2.4. In this section, we express the model properties of h using its Fourier coefficients. The detailed proofs are deferred to Appendix C.

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 2.1). As we focus on the Boolean case, the worst case perturbation Γ_ρ is the protocol of flipping vector coordinates with a probability ρ. Specifically, a corrupted sample $\mathbf{y}$ is generated from $\mathbf{x}$ such that for every component, we independently set $y_i = x_i$ with probability $\frac{1+\rho}{2}$ and $y_i = -x_i$ with probability $\frac{1-\rho}{2}$. This perturbation mechanism leads us to the ρ-flipping influence function.

Definition 3.3 (ρ-flipping Influence Function)**.** The ρ-flipping influence function of any model h is defined as $\mathrm{Inf}_\rho(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim\Gamma_\rho(\mathbf{x})}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

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

Proposition 3.4. *Robustness of h under the Γ_ρ flipping perturbation is equivalent to the ρ-flipping influence function, and thus, can be expressed as*

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

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

Definition 3.5 ((ρ, l)-flipping influence function)**.** The (ρ, l)-flipping influence function of any model h is defined as $\mathrm{Inf}_{\rho,l}(h) = \mathbb{P}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim N_{\rho,l}(\mathbf{x})}[h(\mathbf{x}) \neq h(\mathbf{y})]$.

We leverage (ρ, l)-flipping influence function to express IF of h in terms of its Fourier coefficients (Proposition 3.6).

Proposition 3.6. *Individual fairness defined with respect to the $\Gamma_{\rho,l}$ perturbation is equivalent to the (ρ, l)-flipping influence function, and thus, can be expressed as*

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

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

Unifying robustness and IF using a characteristic function: It is worth noting that IF is similar to robustness,

differing only by a single degree of freedom, i.e. the number of flipped directions l. Specifically, from Equation (2) and (3), we observe that both the properties as $\mu(h) = \sum_{S\subseteq[n]} \texttt{char}(S,\mu_.)\hat{h}(S)^2$, such that $\texttt{char}(S,\mu_{\text{Rob}}) = \rho^{|S|}$, and $\texttt{char}(S,\mu_{\text{IFair}}) = \rho^{|S_l|}$. We call char as *the characteristic function of the property*.

c. Group Fairness (GF): Now, we focus on Group Fairness which aims to ensure similar predictions for different subgroups of population (Barocas, Hardt, and Narayanan 2023). We focus on Statistical Parity (SP) as the measure of deviation from GF (Feldman et al. 2015). To quantify SP, we propose a novel membership influence function.

Definition 3.7 (Membership influence function)**.** If A denotes a sensitive feature, we define the membership influence function w.r.t. A as the conditional probability $\text{Inf}_A(h) \triangleq \mathbb{P}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}}\Big[h(\mathbf{x}) \neq h(\mathbf{y})\Big| x_A = 1, y_A = -1\Big]$.

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

Note that the membership influence function is a randomized version of the deterministic influence function in (O'Donnell 2014). If we denote the transformation of the flipping membership, i.e., the sensitive attribute of $\mathbf{x}$, $f_A(\mathbf{x})$, the classical influence function is $\text{Inf}_A^{det} = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) \neq h(f_A(\mathbf{x}))]$. 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 of $\mathcal{D}$. Thus, it fails to encode the information relevant to SP, whereas the proposed membership influence function does it correctly as shown below.

Proposition 3.8. *Statistical parity of h with respect to a sensitive attribute A and distribution $\mathcal{D}$ is the root of the second order polynomial $P_{\hat{h}}(X)$, i.e. $\alpha(1-\alpha)X^2 - \hat{h}(\emptyset)(1-2\alpha)X - \sum_{S\subseteq[n],S\ni A}\hat{h}(S)^2 - \frac{(1-\hat{h}^2(\emptyset))}{2}$, where $\alpha = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[x_A = 1]$ and $\hat{h}(\emptyset)$ is the coefficient of empty set.*

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

3.3 NP-hardness of Exact Computation

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

Algorithm 1: Active Fourier Auditor (AFA)

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

Theorem 3.9. *Let $\mathcal{Q} \triangleq \{\mathbf{x}, h(\mathbf{x})\}$ be the set of input samples sent to h and the predictions obtained. Given $\tau \in \mathbb{R}_{\geq 0}$, exactly computing all the τ-significant Fourier coefficients of h is NP-complete.*

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

3.4 Algorithm: Active Fourier Auditor (AFA)

We have shown that finding significant Fourier coefficients can be an NP-hard problem. In this section, we propose AFA (Algorithm 1) that takes as input a *restricted access* of $q > 0$ queries from the data-generating distribution and requests labels from the black-box oracle of h (Line 2). Those queries enable us to find the squares of significant Fourier coefficients and estimate them simultaneously. The list of the significant Fourier coefficients L_h of the model h contains both subsets and their estimated Fourier weights. We adopt a Goldreich-Levin (GL) algorithm based approach (Goldreich and Levin 1989; Kushilevitz and Mansour 1993) to find such list of significant Fourier coefficients (Appendix A). Since estimating the properties – robustness, individual fairness, and group fairness – depend on estimating those Fourier coefficients, we plug in their computed estimates and output an (ϵ,δ)-PAC estimate of the properties (Line 4 and 5).

Algorithmic Insights: To compute the significant Fourier coefficients, we start with the power set. Now, we denote the subsets containing an element i as $\mathcal{B}_i(\mathcal{X})$, and the subsets not containing i as $\mathcal{B}_{\neg i}(\mathcal{X})$. Let Υ denote a trajectory starting from the set of all Fourier coefficients in the binary

search tree of Fourier coefficients (Figure 2). The question is that *from the power set, how can we design a Υ to reach subsets of Fourier coefficients above a given threshold τ?*

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

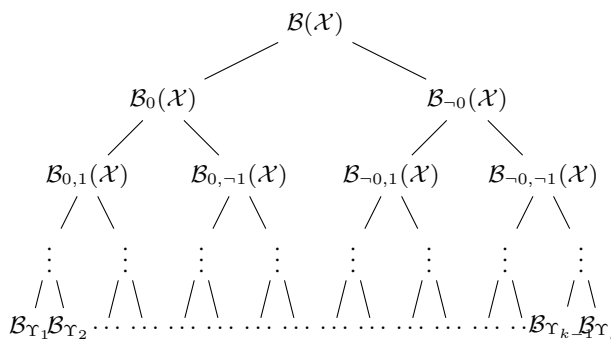

Figure 2: AFA begins with the set of all Fourier coefficients, with weight equal to one ($\tau < 1$). It proceeds by splitting the bucket and verifies at each level of the tree the weight of the node. If the weight is below the threshold, the algorithm halts. Otherwise, it continues to expand, yielding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

Extension to Continuous Features: (Heidari et al. 2021) extend Proposition 2.4 to encompass a general Euclidean space. We use the generic construction of Fourier coefficients in the Euclidean space to extend our computations for feature spaces involving both categorical and continuous features. Rest of our computations follow naturally.

Extension to Multi-class Classification: We also deploy AFA for multi-class classification, where $\mathcal{Y}$ consists of multiple labels. In this setting, the concept of group fairness, i.e. $\mu_{\text{GFair}}(h) \triangleq \max_{y \in \mathcal{Y}} |\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x}) = y | x_A = 1] - \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x}) = y | x_A = -1]|$, is called multicalibration (Dwork et al. 2023). Here, we construct Fourier expansions of the model for each pair of labels. Then, we use Proposition 3.8 to compute the group fairness for each of the expansions, and finally, take the maximum to estimate multi-group fairness of h. Formal details are deferred to Appendix E. We experimentally evaluate both the extensions.

4 Theoretical Analysis

In this section, we establish theoretical guarantees for our proposed auditing framework. First, we prove that AFA is manipulation-proof with respect to statistical parity, demonstrating invariance of the estimation error bounds under a broad class of strategic gaming of the (pre-audit) decision rule. Subsequently, we derive information-theoretic lower bounds to quantify the cost of manipulation robustness thereby establishing optimality criteria for our approach.

Theorem 4.1 (Upper bounds for Robustness and Individual Fairness)**.** AFA *is a PAC-agnostic auditor for robustness and individual fairness with sample complexity* $\mathcal{O}\Big(\frac{\texttt{char}(L,\mu)(1-4\texttt{char}(\bar{L},\mu))}{\epsilon}\sqrt{\log\frac{2}{\delta}}\Big)$. *Here,* $\texttt{char}(L, \mu.) \triangleq \sum_{S \in L} \texttt{char}(S, \mu.)$ *and* $\texttt{char}(\bar{L}, \mu.) \triangleq \sum_{S \in \bar{L}} \texttt{char}(S, \mu.)$.

Theorem 4.2 (Upper bounds for Group Fairness)**.** AFA *yields an (ϵ, δ)-PAC estimate of $\mu_{\text{GFair}}(h)$ if it has access to predictions of $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{4}{\delta}\right)$ input samples.*

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

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

Definition 4.3 (Fourier strategic manipulation-proof)**.** Let h be a model that admits a Fourier expansion as in $h = \sum_{S \subseteq [n]} \hat{h}(S)\psi_S$. We say that an auditor $\mathcal{A}$ achieves optimal manipulation-proofness for estimating a (distributional) property μ when $\mathcal{A}$ is a PAC-agnostic auditor (Definition 3.1) and outputs an exponential-size subclass of functions that satisfies $\forall h, h' \in \mathcal{M}, \mathbb{P}(|\mu(h) - \mu(h')| \geq \epsilon) \leq \delta$.

Theorem 4.4 (Manipulation-proofness of AFA)**.** AFA *achieves optimal manipulation-proofness for estimating statistical parity with manipulation-proof subclass of size 2^{n-2}.*

Lower Bounds without Manipulation-proofness: THeorem 4.5 gives a lower bound for yielding a PAC estimate of the statistical parity without manipulation-proof. The proof is in Appendix D.3.

Theorem 4.5 (Lower bound without manipulation-proofness)**.** *Let $\epsilon \in (0, 1)$, $\delta \in (0, 1/2]$. We aim to obtain (ϵ, δ)-PAC estimate of SP of model $h \in \mathcal{H}$, where the hypothesis*

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

Table 1: Average estimation error for statistical parity across different ML models. '—' denotes when a method cannot scale to the model. The best method is in **bold**.

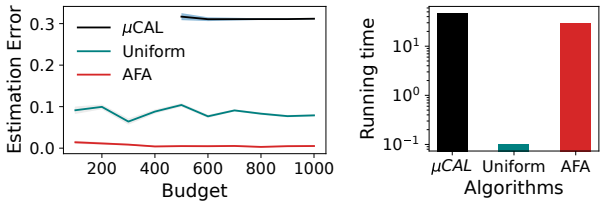

Figure 3: Error (left) and running time (right) of different auditors in estimating statistical parity of COMPAS in LR.

class $\mathcal{H}$ has VC dimension d. For any auditing algorithm $\mathcal{A}$, there exists an adversarial distribution realizable by the model to audit such that with $\tilde{\Omega}(\frac{\delta}{\epsilon^2})$ samples, $\mathcal{A}$ outputs an estimate $\hat{\mu}$ of $\mu_{\text{GFair}}(h^*)$ with $\mathbb{P}[|\hat{\mu} - \mu_{\text{GFair}}(h^*)| > \epsilon] > \delta$.

Our results extend the existing sample complexity results with model reconstruction (Yan and Zhang 2022), and also provide a reference of optimality for upper bounds. We highlight the gap from the upper bound established in Theorem 4.2, attributed to the lack of the manipulation proof.

5 Empirical Performance Analysis

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

Experimental Setup: We conduct experiments on COMPAS (Angwin et al. 2016), student performance (Student) (Cortez and Silva 2008), and drug consumption (Drug) (Fehrman et al. 2019) datasets. The datasets contain a mix of binary, categorical, and continuous features for binary and multi-class classification. We evaluate AFA on three ML models: Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF). The ground truth of group fairness, individual fairness, and robustness is computed using the entire dataset as in (Yan and Zhang 2022). For group fairness, we compare AFA with uniform sampling method, namely `Uniform`, and the active fairness auditing algorithms (Yan and Zhang 2022, Algorithm 3), i.e. `CAL` and its variants μ`CAL` and randomized μ`CAL`, which requires more information about the model class than black-box access. We report the best variant of `CAL` with the lowest error. For robustness and individual fairness, we compare AFA with `Uniform`. Each experiment is run 10 times and we report the averages. We refer to Appendix F.1 for details.

Our empirical studies have the following **objectives**:

1. How accurate AFA is with respect to the baselines to audit robustness, individual fairness, and group fairness for different models and datasets?
2. How sample efficient and computationally efficient AFA is with baselines in auditing distributional properties?

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

Table 2: Estimation error for robustness and individual fairness by `Uniform` and AFA. **Bold** case means lower error.

Accurate, Sample Efficient, and Fast Estimation of Group Fairness: In Table 1, we demonstrate the estimation error of group fairness by different methods across datasets and models. AFA yields the lowest estimation error, hence a better method, than all baselines in all nine configurations of models and datasets. Among baselines, `CAL` cannot estimate group fairness beyond COMPAS on LR, due to the requirement of a finite version space, which is provided only for COMPAS on LR. `Uniform`, albeit simple to implement, invariably demonstrates erroneous estimate. *Thus,* AFA *is the most accurate auditor for group fairness w.r.t. baselines.*

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

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

6 Conclusion and Future Work

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

Acknowledgements

This work is supported by the Regalia project of Inria and French Ministry. D. Basu further acknowledges the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), the PEPR project FOUNDRY (ANR23-PEIA-0003), and the CHIST-ERA project CausalXRL (ANR-21-CHR4-0007) for their support.

References

Albarghouthi, A.; D'Antoni, L.; Drews, S.; and Nori, A. V. 2017. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA): 1–30.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

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

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Cohn, D.; Atlas, L.; and Richard, L. 1994. Improving generalization with active learning. In *Machine Learning (20)*, 201–221.

Cortez, P.; and Silva, A. M. G. 2008. Using data mining to predict secondary school student performance.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 12)*, volume 86, 214—226.

Dwork, C.; Lee, D.; Lin, H.; and Tankala, P. 2023. From pseudorandomness to multi-group fairness and back. In *The Thirty Sixth Annual Conference on Learning Theory*, 3566–3614. PMLR.

Fehrman, E.; Egan, V.; Gorban, A. N.; Levesley, J.; Mirkes, E. M.; and Muhammad, A. K. 2019. *Personality traits and drug consumption*. Springer.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Friedler, S.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im)possibility of fairness. In *arxiv*. Arxiv:1609.07236.

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

Ghosh, B.; Basu, D.; and Meel, K. S. 2021. Justicia: A stochastic SAT approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7554–7563.

Ghosh, B.; Basu, D.; and Meel, K. S. 2022. Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9539–9548.

Goldreich, O.; and Levin, L. 1989. 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*.

Goldwasser, S.; N. Rothblum, G.; Shafer, J.; and Yehudayoff, A. 2021. Interactive Proofs for Verifying Machine Learning. *Innovations in Theoretical Computer Science Conference (ITCS)*.

Heidari, M.; Sreedharan, J.; Shamir, G. I.; and Szpankowski, W. 2021. Finding Relevant Information via a Discrete Fourier Expansion. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139:4181-4191*.

Herman, T.; and Rothblum, G. N. 2022. Verifying the Unseen: Interactive Proofs for Label-Invariant Distribution Properties. *STOC: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*.

Ilyas, A.; Min Park, S.; Engstrom, L.; Leclerc, G.; and Madry, A. 2022. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

Jagielski, M.; Carlini, N.; Berthelot, D.; Kurakin, A.; and Papernot, N. 2020. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, 1345–1362.

John, P. G.; Vijaykeerthy, D.; and Saha, D. 2020. Verifying Individual Fairness in Machine Learning Models. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Kearns, M.; Neel, S.; Roth, A.; and Steven Wu, Z. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *Proceedings of the 35th International Conference on Machine Learning, PMLR*.

Kumar, R. S. S.; Nyström, M.; Lambert, J.; Marshall, A.; Goertzel, M.; Comissoneru, A.; Swann, M.; and Xia, S. 2020. Adversarial machine learning-industry perspectives. In *2020 IEEE security and privacy workshops (SPW)*, 69–75. IEEE.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2016. Adversarial Machine Learning at Scale. In *International Conference on Learning Representations*.

Kushilevitz, E.; and Mansour, Y. 1993. Learning decision trees using the Fourier spectrum. *SIAM J. Comput., 22(6):1331–1348*.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

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

Mathieu, T.; Basu, D.; and Maillard, O.-A. 2022. Bandits Corrupted by Nature: Lower Bounds on Regret and Robust Optimistic Algorithms. *Transactions on Machine Learning Research*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; Sandvig, C.; et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

Mutreja, S.; and Shafer, J. 2023. PAC Verification of Statistical Algorithms. *36th Annual Conference on Learning Theory (COLT)*.

Neiswanger, W.; Wang, K. A.; and Ermon, S. 2021. Bayesian Algorithm Execution: Estimating Computable Properties of Black-box Functions Using Mutual Information. *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

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

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. 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*, 33–44.

Rasheed, K.; Qayyum, A.; Ghaly, M.; Al-Fuqaha, A.; Razi, A.; and Qadir, J. 2022. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 149: 106043.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32.

Wang, Y.; Baharav, T. Z.; Han, Y.; Jiao, J.; and Tse, D. 2022. Beyond the Best: Estimating Distribution Functionals in Infinite-Armed Bandits. *arXiv preprint arXiv:2211.01743*.

Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677.

Xu, H.; and Shie, M. 2011. Robustness and generalization. In *Maéch Learn*, volume 86, 391—423.

Yan, T.; and Zhang, C. 2022. Active fairness auditing. In *International Conference on Machine Learning*, 24929–24962. PMLR.