EFFICIENT ACTIVE AUDITING OF MULTI-GROUP FAIRNESS WITH BIAS PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Over the past decade, Machine Learning (ML) has been trained under dual objectives: minimizing prediction error via Empirical Risk Minimization (ERM) while controlling unfairness bias. In practice, however, fairness-aware training often yields limited improvements over standard ERM, making reliable post hoc auditing essential. Existing auditing approaches for black-box models either rely on model reconstruction –exposing systems to extraction attacks– or directly estimate fairness metrics, offering limited insight into which regions of the data distribution drive bias. More fundamentally, property-specific auditing –aimed at extracting only targeted fairness information without reconstructing the model– remains poorly understood. In this work, we introduce the bias probe framework, which enables targeted and adaptive querying to reveal bias structure while preserving model confidentiality. Building on this framework, we propose ALeBi, an active auditor that learns such probes to efficiently estimate multi-group fairness metrics. We establish novel sample complexity guarantees governed by a property-specific complexity measure, resolving a previously posed open question, and extend our analysis to adversarial settings where the model owner may strategically obscure bias. Our results uncover a fundamental trade-off between model confidentiality and reliable auditing, and show that property-specific probing enables both accurate estimation and interpretable identification of high and low-bias regions. Extensive experiments support our theoretical findings and demonstrate the practical effectiveness of our approach.

1 INTRODUCTION

Assessing ML models has become increasingly critical, as mounting evidence shows that deployed systems may exhibit harmful biases across a wide range of domains, including medical imaging (Seyyed-Kalantari et al., 2021; Daneshjou et al., 2022; Vrudhula et al., 2024; Banerjee et al., 2021), hiring and job screening (Harwell, 2022; Wilson & Caliskan, 2024), credit lending (Garcia et al., 2024), and online platforms (Biswas & Rajan, 2020). Among these concerns, discriminatory bias has received particular attention, giving rise to the field of algorithmic fairness (Barocas & Selbst, 2016; Hardt et al., 2016). A central line of work focuses on designing fair learning procedures, where fairness constraints are incorporated directly into the training process, typically as optimization constraints trading off predictive accuracy and fairness guarantees (Zafar et al., 2017; Agarwal et al., 2018). However, empirical evidence suggests that such approaches may yield limited improvements over standard ERM in practice (Zong et al., 2022).

A large body of work has studied fairness auditing under this paradigm. Existing approaches typically fall into two categories. The first relies on *model reconstruction*, using tools from active learning (Angluin, 1988; Balcan et al., 2012) to learn the model –or an accurate surrogate– before estimating fairness metrics. The second relies on *direct estimation* of a difference of means, using labeled samples. However, both approaches exceed the auditing objective: reconstruction attempts to recover the full model, while direct estimation is as costly as replicating the learning process and do not produce explanations, making it impractical in realistic auditing scenarios. Moreover, reconstruction-based approaches may expose the model to extraction attacks, compromising confidentiality. This tension raises the following question:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

> ***Q1***: Can we obtain distribution-free guaranties for fairness auditing that avoid both full model reconstruction and large-scale labeled data requirements?

Beyond estimation, interpretability is also a critical component of auditing. In the context of machine learning regulation, Article 86 of the AI Act European Commission (2024b) underscores the right to explanation in automated decision-making. For example, an individual denied a bank loan by an ML system is entitled to receive an explanation for that decision. While standard audit methods provide quantitative summaries of bias (Goldreich et al., 1998; Yan & Zhang, 2022; Chugg et al., 2023), they do not yield interpretable representations of the model's discriminatory behavior. In parallel, the literature on representation learning has introduced *probes* as simple functions used to extract specific properties from learned representations (Alain & Bengio, 2016). Inspired by this perspective, we introduce *bias probes* – structured comparison functionals that capture relational disparities between protected groups. Unlike feature attribution methods, which operate at the level of individual inputs, bias probes focus on intergroup comparisons and provide an interpretable summary of discriminatory behavior. To the best of our knowledge, this is the first work to formalize fairness auditing with explicit interpretability guarantees in this sense.

Finally, we must account for the security of the auditing process. Existing auditing methods, particularly those that rely on membership queries (Yan & Zhang, 2022), may render the model vulnerable to extraction attacks, effectively reducing auditing to model reconstruction. This highlights a second fundamental challenge: ensuring that auditing procedures do not compromise model confidentiality. This leads to our second question:

> ***Q2***: Can we design sample-efficient and interpretable auditing procedures while preserving model confidentiality against extraction attacks?

Beyond the accuracy and interpretability of the audit report, we further extend our setting by considering a model owner who may behave adversarially, with the goal of concealing unfairness across groups. The adversarial behavior we consider differs from the adversarial models commonly studied in the literature, such as Huber contamination models Chen et al. (2016), adversarial distribution shifts Croce et al. (2020), or adversarial input perturbations Montasser et al. (2021); Goodfellow et al. (2018). In these settings, the adversary typically manipulates the data or the input distribution, whereas our adversary is *fairness-aware*: the model owner may strategically manipulate the model's behavior to make unfairness harder to detect by the auditor. This leads to our second research question:

> ***Q3***: How can we design fairness auditors that are accurate and interpretable while remaining robust to fairness-aware adversaries?

Related works. (i) Fairness Auditing: Algorithmic fairness has been widely studied from both the learning and auditing perspectives. A large body of work focuses on learning fair models via constrained optimization or reduction-based approaches (Zafar et al., 2017; Agarwal et al., 2018). In addition, post hoc audit methods aim to assess the fairness properties of a model deployed under limited access (Kearns et al., 2018; Yan & Zhang, 2022; Chugg et al., 2023). Recent works have proposed active auditing strategies for estimating fairness metrics under distributional assumptions (Yan & Zhang, 2022). These approaches often reduce auditing to model reconstruction within a hypothesis class, leading to guarantees that scale with the complexity of learning. Other related works propose frameworks, where bias audit reduces to a binary objective via testing approaches (Chugg et al., 2023; Goldreich et al., 1998). In contrast, our work treats fairness auditing as a property estimation problem, focusing on the intrinsic combinatorial structure of intergroup comparisons rather than recovering the full model. (ii) Interpretability and Probing Methods: Post hoc interpretability methods, including feature attribution techniques (Ribeiro et al., 2016; Lundberg & Lee, 2017) and counterfactual explanations (Kusner et al., 2017), aim to provide descriptive summaries of model behavior. In parallel, probing methods have been developed in representation learning to extract specific properties from learned representations using simple models (Vafa et al., 2025; Alain & Bengio, 2016) or alternatively explaining biases in Large Language Models Immer et al. (2022); Morehouse et al. (2025); Guo et al. (2022); Manerba et al. (2024). Our notion of bias probes is conceptually related but differs in both objective and scope. Rather than analyzing internal representations, bias probes capture relational disparities between protected groups and are directly

tied to fairness properties. Moreover, our probes are equipped with statistical guarantees, bridging interpretability and rigorous auditing.

Summary of contributions. Our main contributions are as follows:

(i) **Comparison-based auditing.** We introduce *bias probes*, formalizing multi-group fairness auditing as learning structured comparison functionals with distribution-free guarantees.

(ii) **Query complexity.** We introduce the *k-inter-group star number* $\mathfrak{s}_k$ and characterize auditing complexity through upper and lower bounds, with k-ALEBI achieving the upper bound.

(iii) **Model-extraction hardness.** We establish computational hardness of identifying the underlying model from relational queries, showing that accurate fairness auditing does not require model recovery.

(iv) **Robust auditing.** We introduce a fairness-aware manipulation model and ROBUST k-ALEBI, which remains robust under conditionally bounded corruption with $p < 1/2$.

(v) **Empirical evaluation.** We evaluate accuracy, interpretability, model-extraction resistance, and robustness to manipulation against RECONAUDIT, DIRECTAUDIT, and Yan & Zhang (2022).

2 PROBLEM FORMULATION: AUDITING MULTI-GROUP FAIRNESS

Notation. Let $\mathcal{H}$ denote a hypothesis class. The model owner $\mathbf{M}$ trains a model $h : \mathcal{Z} \subseteq \mathbb{R}^{p+1} \to \tilde{\mathcal{Y}}$ from $\mathcal{H}$ on a dataset which then deploys. We assume that the first p coordinates correspond to unprotected features, while the $(p+1)$-th coordinate represents a protected attribute A. This attribute partitions the input space $\mathcal{Z}$ into $k \geq 2$ disjoint subpopulations $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$ (e.g., when A represents gender, $k = 2$). We denote by $\mathbb{1}_{[\cdot]}$ the indicator function. Finally for $a \in \mathbb{R}$, $[a]_+ := \max(a, 0)$.

We study group-level properties of h. In particular, we consider the maximum disparity of h across protected groups, captured by multi-group statistical parity.

Definition 1 (Multi-Group Fairness). *Let $h : \mathcal{Z} \to \tilde{\mathcal{Y}}$ and let $\mathcal{D}$ be a distribution over $\mathcal{Z}$. For a pair of groups $(i, j) \in [k]^2$, define $\mu_{i,j}(\mathcal{D}, h) = \mathbb{E}_{X \sim \mathcal{D}}[h(X) \mid X \in \mathcal{Z}_i] - \mathbb{E}_{X \sim \mathcal{D}}[h(X) \mid X \in \mathcal{Z}_j]$. The multi-group fairness violation of h is $\mu(\mathcal{D}, h) = \max_{i,j \in [k]} \mu_{i,j}(\mathcal{D}, h)$.*

The functional $\mu(\mathcal{D}, h)$ depends jointly on the model and the data distribution. In particular, disparities in regions of $\mathcal{Z}$ with negligible probability mass under $\mathcal{D}$ do not contribute substantially to μ. Therefore, as illustrated in Figure 3, estimating μ requires access to the marginal distribution over $\mathcal{Z}$, which motivates an active audit setting in which information about $\mathcal{D}$ is obtained through a pool of unlabeled points $\mathcal{P}$.

More generally, we consider an auditing setting in which an auditor $\mathbf{A}$ has access to a pool of unlabeled samples $\mathcal{Q}$ and interacts with a model owner $\mathbf{M}$ by issuing queries $q_1, \ldots, q_m$ in order to estimate a property μ of a model h. Formally, $\mu : \mathcal{P}(\mathcal{Z}) \times \mathcal{H} \to \mathbb{R}$ is a functional that quantifies a distributional property of models in $\mathcal{H}$. An auditing problem is defined by the tuple $\langle \mathcal{Z}, \tilde{\mathcal{Y}}, \mathcal{P}, \mathrm{I}_{\mathbf{M}}(h), \mu, \ell_\mu \rangle$, where $\mathcal{Z}$ and $\tilde{\mathcal{Y}}$ denote the input and output spaces, $\mathcal{P}$ is a class of probability measures over $\mathcal{Z}$, and $\mathrm{I}_{\mathbf{M}}(h)$ represents the information about h available to the auditor. In the white-box setting, $\mathrm{I}_{\mathbf{M}}(h) = h$, while in the black-box setting it corresponds to oracle access to h, possibly augmented with side information such as the hypothesis class $\mathcal{H}$. The functional μ captures a property of h under distributions in $\mathcal{P}$, and $\ell_\mu : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a bounded loss function. Additional details about the general auditing framework are provided in Appendix A. We now introduce the notion of PAC auditing in the interactive setting. We focus on auditing multi-group properties, which introduce additional challenges as reliable estimation requires uniform information across subpopulations. In particular, we study the auditing of multi-group fairness (Definition 1). To date, several approaches have been proposed to audit statistical parity. Broadly, these methods fall into two categories: Audits via model reconstruction and Audits via direct Estimation. Details, including interaction protocols and pseudocodes are given in Appendix B.

3 INTERPRETABLE, MANIPULATION-PROOF AND RECONSTRUCTION-FREE AUDITS WITH BIAS PROBES

In this section, we introduce the auditing setting that will be used throughout the remainder of the paper. Rather than extracting audit-relevant information directly from the input space $\mathcal{Z}$, we consider a conceptual framework in which such information is extracted over the product space of protected groups. This perspective allows us to design an auditing framework that focuses exclusively on the fairness property under audit, while providing interpretable audit information and mitigating the risk of model extraction.

Lifting the Sample Pool to the Product Space. Let $\mathcal{X} \triangleq \prod_{i=1}^{k} \mathcal{Z}_i$ denote the product space induced by the k protected subpopulations. An element $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{X}$ consists of one sample from each protected group, and $\mathcal{X}$ serves as the query space for auditing multi-group fairness (Definition 1). This perspective motivates a new auditing formulation in which the relevant information is represented by a class of functions defined directly on the query space, as introduced below:

Definition 2 (Multi-Context Class)**.** *For a hypothesis class $\mathcal{H}$. The relative multi-context class induced by $\mathcal{H}$ is defined as:*

$$\mathcal{F}_k(\mathcal{H}) = \left\{ f_h : \mathcal{X} \to \mathbb{R} \;\middle|\; f_h(\mathbf{x}) = \mathbf{M}_k\big(h(x_1), \ldots, h(x_k)\big),\ h \in \mathcal{H} \right\},$$

where $\mathbf{M}_k$ is a fixed operator that depends on the property under audit.

This perspective naturally induces an interaction model between the auditor $\mathbf{A}$ and the model owner $\mathbf{M}$. Unlike standard membership-query access, the auditor cannot query the model on individual points or observe its absolute predictions. Instead, the auditor submits a structured query $\mathbf{X} \in \mathcal{X}$, consisting of one sample from each protected group. The auditor's access to the model is mediated by the oracle $\mathrm{I}_{\mathbf{M}}(\mathbf{X}) = \mathbf{M}_k(\mathbf{X})$, which, for our multi-group fairness property, returns only comparative evaluations across groups: $\mathbf{M}_k(\mathbf{X}) = \big(f_{i,j}(\mathbf{X})\big)_{i \neq j}$. Thus, the auditor observes relative disparities between groups rather than the model's absolute predictions.

Definition 3 (Cross-Group Queries (CGQ))**.** *Given access to a pool of unlabeled samples, a Cross-Group Query consists of selecting a tuple $\mathbf{X} \in \mathcal{X}$ formed by choosing one sample from each protected group. The response to $\mathbf{X}$ is the collection of pairwise comparisons $\big(f_{i,j}(\mathbf{X})\big)_{i \neq j}$.*

Kane et al. (2017) proposed a related type of query in the context of active classification with linear separators. Their comparison queries consist of presenting a pair of points to an oracle, which indicates which point lies closer to the decision boundary. In our setting, instead, the auditor presents a pair of points from different groups to the model owner, who identifies the pair exhibiting the largest fairness disparity.

Coupling and Sampling Invariance. We now formalize the role of the operator $\mathbf{M}_k$ in Definition 2, in other words, the role of sampling strategies over the product space from the pool $\mathcal{Q}$. Let ν be any probability measure on $\mathcal{X}$ whose marginals satisfy $\nu_i(\cdot) = \mathcal{D}(\cdot \mid \mathcal{Z}_i)$ for every $i \in [k]$. Let $\nu_{i,j}$ denote the corresponding pairwise marginal on $\mathcal{Z}_i \times \mathcal{Z}_j$. The following result shows that the choice of coupling does not affect the value of μ.

Proposition 1 (Coupling Invariance)**.** *Let μ denotes multi-group fairness defined in 1.*

Let $\nu \in \Pi(\nu_1, \ldots, \nu_k)$ be any coupling measure on $\mathcal{X}$ whose ith marginal satisfies $\nu_i(\cdot) = \mathcal{D}(\cdot \mid \mathcal{Z}_i)$, for all $i \in [k]$. The following holds:

$$\mu(h, \mathcal{D}) = \left\| \mathbb{E}_{\mathbf{X} \sim \nu}[f_h(\mathbf{X})] \right\|_{\infty}$$

This invariance implies that auditing multi-group fairness can be carried out without modeling the joint distribution across groups, and justifies the use of arbitrary coupling strategies when constructing queries from the sample pool. The proof is given in Appenidx C.

3.1 INTERPRETABLE AUDITS VIA BIAS PROBES

To obtain interpretable insights into the model's behavior, we introduce bias probes, which localize the comparative behavior of h across protected groups.

Definition 4 (Bias Probes). *Let $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{X}$. **(i) Global bias probe.** The global bias probe is the function $f : \mathbf{x} \mapsto \big(f_{i,j}(\mathbf{x})\big)_{i \neq j}$. **(ii) Feature-wise bias profile.** For $q \in [p]$, let $P_q(\mathbf{X}) = \big(X_1^{(q)}, \ldots, X_k^{(q)}\big)$ denote the projection of the CGQ query onto the qth unprotected feature. For each pair (i, j), define the conditional feature-wise probe $\mathfrak{e}_{q,i,j}(z) := \mathbb{E}_{\mathbf{X} \sim \mathcal{D}^k}\left[f_{i,j}(\mathbf{X}) \mid P_q(\mathbf{X}) = z\right]$. The vector $\mathfrak{e}_q(z) = \big(\mathfrak{e}_{q,i,j}(z)\big)_{i<j}$ describes how the expected relational response of the deployed model varies with feature q. **(iii) Feature-wise bias magnitude.** We summarize the conditional relational disparity by $\widetilde{f}_q(z) := \max_{i<j} |\mathfrak{e}_{q,i,j}(z)| = \max_{i<j} |\mathbb{E}\left[f_{i,j}(\mathbf{X}) \mid P_q(\mathbf{X}) = z\right]|$.*

An example of application is given in Appendix A.2. Together, these three quantities provide complementary levels of evidence: the global probe establishes whether the model exhibits discrimination overall, the feature-wise probes indicate *where* disparities vary across the input space, and the expected feature-wise probes identify regions in which the disparities are most pronounced. The resulting audit report therefore provides the model owner not only with evidence of unfairness, but also with interpretable information about where and along which features that unfairness manifests itself. We now formalize interpretable active auditing for general properties μ through the lens of comparison function learning.

Definition 5 (Interpretable Active Auditing). *Let $\mathcal{P}$ be a class of distributions over $\mathcal{Z}$, and let $\mathcal{F}$ be an intergroup context class defined over $\mathcal{X}$. We say that $\mathcal{F}$ is actively learnable under $\mathcal{P}$ if there exists a sample complexity function $m_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$, such that for all $(\varepsilon, \delta) \in (0,1)^2$, for all $\mathcal{D} \in \mathcal{P}$, and for all $m \geq m_{\mathcal{F}}(\varepsilon, \delta)$, there exists an algorithm $\mathcal{A}_m$ which, based on m queries $\mathcal{Q}_1, \ldots, \mathcal{Q}_m$, outputs an estimate $\hat{f}_m \in \mathcal{F}$ satisfying $\mathbb{P}_{\mathcal{Q}_1, \cdots \mathcal{Q}_m}\Big[\mathcal{E}_{\mathcal{D}}(\hat{f}_m) - \inf_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{D}}(f) > \varepsilon\Big] < \delta$, where the risk $\mathcal{E}_{\mathcal{D}}(f)$ is defined as $\mathcal{E}_{\mathcal{D}}(f) := \mathbb{P}_{\mathbf{X} \sim \mathcal{D}^k}\big[f(\mathbf{X}) \neq f_h(\mathbf{X})\big]$, and f_h denotes the comparison functional induced by the model h. The minimax sample complexity of the auditing problem is defined as $\inf_{m_{\mathcal{F}}} \sup_{\mathcal{D} \in \mathcal{P}} m_{\mathcal{F}}(\varepsilon, \delta)$, where the infimum is taken over all valid active auditing strategies.*

This formulation reduces the auditing problem for μ to the problem of learning the comparison functional f_h, from which $\mu(\mathcal{D}, h)$ can be estimated as a functional of f_h.

3.2 Information Complexities For Learning Probes

In the following section, we define the information-complexities relevant for characterizing the auditing problem of multi-group fairness. We begin with Natarajan dimension Daniely et al. (2015), that will serve to characterize global probe learning in the passive setting:

Definition 6 (k-Daniely-Shwartz Dimension (Daniely et al., 2015)).

- ***Pseudo-cube.*** *Let $B \subseteq \mathcal{Y}^m$ be finite and non-empty. Two vectors $b, b' \in B$ are called i-neighbors if $b_i \neq b'_i$ and for every $j \neq i$, $b_j = b'_j$. The set B is an m-dimensional pseudo-cube if for every $b \in B$ and every $i \in [m]$, b has an i-neighbor in B.*
- ***Daniely-Shwartz dimension.*** *A sequence $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathcal{X}_k^m$ is DS-shattered by the k-context class $\mathcal{F}_k$ if the set $\mathcal{F}_k|_S = \{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)) : f \in \mathcal{F}_k\}$ contains an m-dimensional pseudo-cube.*

  Define Daniely-Shwartz dimension $\mathrm{DS}_k(\mathcal{F})$ as the largest DS-shattered dimension, or ∞ if no finite largest dimension exists.

Definition 7 (Multi-star Number). *Let $\mathcal{F}$ be a class of bias probes defined on $\mathcal{X}$.*

The multi-star number $\mathfrak{s}_k$ *is the largest value of m for which there exist distinct points $\{\mathbf{X}_i : i \in [m]\}$, and functions $f_0 \in \mathcal{F}, \quad f_i \in \mathcal{F}$, such that for all $i \in [m]$, $\mathrm{DIS}(\{f_0, f_i\}) = \{\mathbf{X}_i\}$,* and each component p of the multi-context vector $\mathbf{X}$ belongs to the protected group $\mathcal{Z}_p$, for $p \in [k]$. *If no finite maximum exists, we set $\mathfrak{s}_k = \infty$.*

3.3 Information-Theoretic Separation of Reconstruction and Probe Auditing

It is known that linear classifiers, even in low dimension input space (e.g., $d = 2$), the star number is unbounded, implying that active reconstruction-based auditing cannot improve over passive reconstruction and therefore requires a large number of queries, increasing the risk of model extraction.

Proposition 2 establishes a separation between auditing via active reconstruction and bias probing, the latter avoiding this intrinsic complexity barrier by focusing on property-specific queries.

Proposition 2 (Information-Theoretic Separation of Reconstruction and Probe Auditing)**.** *There exists an audit instance problem, where auditing via model reconstruction is strictly harder than auditing via probe learning.*

This establishes an information-theoretic separation between the two approaches: while reconstruction has unbounded combinatorial complexity for the underlying model class, the corresponding probe-learning problem admits a finite k-multi-star complexity. This aligns with the intuition that multi-group fairness is a global property driven primarily by the protected attribute, rather than the ambient feature space. The proof is given in Appendix B.3.

3.4 Manipulation-free Regime

In this section, we focus on the manipulation-free regime, in which the model owner is honest and does not attempt to conceal unfairness.

3.4.1 Algorithm Design: k-ALeBi and Upper Bounds

Algorithm 1 k-ALeBi

Require: Access to $\mathrm{I}_{\mathbf{M}}(h) = \mathbf{M}_k(h)$, CGQ fixed coupling distribution ν, multi-context probe class $\mathcal{F}$, pool budget n

Initialize version space $\mathcal{V}_0 \leftarrow \mathcal{F}$
for $t = 1$ to n **do**
  Draw independently

$$\mathbf{X}^{(t)} = (X_1^{(t)}, \ldots, X_k^{(t)}) \sim \nu$$

  $\mathcal{V}_t \leftarrow \mathcal{V}_{t-1}$
  if $\mathbf{X}^{(t)} \in \mathrm{DIS}(\mathcal{V}_{t-1})$ **then**
    Query the relational oracle

$$\mathbf{Y}^{(t)} \leftarrow \mathbf{M}_k(h; \mathbf{X}^{(t)}) = f_h(\mathbf{X}^{(t)})$$

$$\triangleright\ \mathbf{Y}^{(t)} \in \{-1, 0, 1\}^{\binom{k}{2}}$$

    Update the version space

$$\mathcal{V}_t \leftarrow \left\{ f \in \mathcal{V}_{t-1} : f(\mathbf{X}^{(t)}) = \mathbf{Y}^{(t)} \right\}$$

  end if
end for
Select a global relational probe

$$\hat{f} \in \mathcal{V}_n$$

Compute

$$\{(f_q, \widetilde{f}_q)\}_{q \in [p]}$$

for all unprotected features q
return

$$\mathsf{AuditReport} = \left(\underbrace{\hat{f}}_{\text{global bias probe}}, \underbrace{\{f_q\}_{q\in[p]}}_{\text{feature-wise bias probes}}, \underbrace{\{\widetilde{f}_q\}_{q\in[p]}}_{\text{expected feature-wise bias probes}} \right)$$

Theorem 3 (Upper bounds of ALeBi)**.** *Algorithm 1 is a PAC active auditor: with probability at least* $1-\delta$*, it outputs* $\hat{f}$ *satisfying* $\mathbb{P}\left\{\hat{f}(Z) \neq f_{h^*}(Z)\right\} \leq \epsilon$ *using at most*

$$\mathcal{O}\left(\min\left\{n,\ (e-1)\min\{\mathfrak{s}_k, n\}\log\frac{en}{\min\{\mathfrak{s}_k,n\}} + \log\frac{2}{\delta} + 1,\ \frac{\min\{\mathfrak{s}_k,n\}}{1\vee\log\min\{\mathfrak{s}_k,n\}}\,\mathrm{DS}_k\log\left(\frac{en(2^k-2)}{\mathrm{DS}_k}\right)\right\}\right)$$

queries. where $n = \left\lceil \frac{\mathrm{DS}_k + \log(4/\delta)}{\epsilon} \right\rceil$.

Corollary 4. *A multi-class* $\mathcal{F}$ *is PAC actively learnable (Definition 5) if and only if* $\mathfrak{s}_k(\mathcal{F}) < \infty$.

Corollary 4 derives from Theorems 14 and 3.

In Section 4, we evaluate the accuracy of k-ALEBI in estimating multi-group fairness (Definition 1) and compare its performance with three baselines: RECONAUDIT, DIRECTAUDIT, and the auditing method proposed by (Yan & Zhang, 2022).

3.5 FAIRNESS-AWARE ADVERSARIAL REGIME

We now consider an adaptive adversarial model owner $\mathbf{M}$ whose decision to attack depends on the auditor's current estimate of multi-group disparity. The robust version of k-ALEBI requires that each CGQ admits a sufficiently homogeneous neighborhood of nonnegligible probability mass. For any CGQ in $\mathcal{Q}$, $J(q)$ denotes a cell from the input space containing q.

Assumption 5 (Local $\mathcal{F}$ replicability)**.** *There exists a measurable neighborhood map* $q \mapsto J(q) \subseteq \mathcal{Q}$ *chosen independently of the unknown target* $f^\star$ *and before observing any* $\mathbf{M}$ *oracle response, such that every CGQ* q *requested by the clean algorithm satisfies a probability mass of:* $\mathbb{P}_Q(J(q)) \geq \gamma$ *and* $\sup_{f \in \mathcal{F}} \mathbb{P}_Q(f(q') \neq f(q) \,|\, q' \in J(q)) \leq \rho$ *for some* $\gamma > 0$ *and* $0 \leq \rho < \frac{1}{2}$.

Therefore a draw from $J(q)$ has the same clean query response as q uniformly over $\mathcal{F}$ with probability at least $1 - \rho$. The cells do not need to form a partition and may overlap. We assume that the learner can sample $q' \sim \mathcal{D}_Q(\cdot \mid J(q))$ from the i.i.d unlabeled pool [1]. Since $\mathbb{P}_Q(J(q)) \geq \gamma$, each accepted local CGQ requires at most $1/\gamma$ unlabeled draws in expectation.

3.5.1 ADAPTIVE FAIRNESS-AWARE ADVERSARY: ALGORITHM AND UPPER BOUNDS

For the r-th local query q'_r, let $A_r \in \{0, 1\}$ denote whether its response is attacked. Let $\mathbb{H}_{r-1}$ denote the complete history before that response. If $A_r = 0$, then $\widetilde{Y}_r = f^\star(q'_r)$. If $A_r = 1$, the adversary may return any $\widetilde{Y}_r \in \mathcal{Y}_k$, possibly as an adaptive function of the history, the queried CGQ q'_r and its clean response.

Assumption 6 (Conditional probabilistic attack)**.** *There exists* $p < 1/2$ *such that*

$$\mathbb{P}(A_r = 1 \,|\, \mathbb{H}_{r-1}, q'_r, f^\star(q'_r)) \leq p$$

for every local query.

In particular, the conditional bound is stronger than the nonadaptive assumption $\mathbb{P}(A_r = 1) \leq p$. We introduce the quantity $\beta \triangleq \rho + (1 - \rho)p = p + \rho - p\rho$, that accounts for both sources of error: the local clean response may differ from the response at q, or a locally correct response may be corrupted. By Assumption 6, $\beta < \frac{1}{2}$.

Let $N_0(\varepsilon, \delta_0)$ denote the upepr bounds derived in Theorem 3. We derive upper bounds for Algorithm **??**.

Theorem 7 (Neighborhood robustification)**.** *Under assumptions 5 and 6, Algorithm 5 is a PAC-active auditor For* $p > 0$*, a sample complexity of* $N_0(\varepsilon, \delta/2) \left\lceil \frac{2}{(1-2\beta)^2} \log \frac{2N_0(\varepsilon,\delta/2)}{\delta} \right\rceil$ *suffices in the presence of fairness-aware adversarial model owner.*

If the neighborhood samples are obtained by rejection sampling from an i.i.d unlabeled stream, the expected number of unlabeled draws is at most $\left\lceil \frac{2N_0(\epsilon,\delta/2)}{\gamma(1-2\beta)^2} \log \frac{2N_0(\varepsilon,\delta/2)}{\delta} \right\rceil$

The pseudocode, the proof and an additional information-theoretic lower bound is given in Appendix F.

[1] this can be implemented via rejection sampling.

4 EXPERIMENTAL ANALYSIS

We empirically evaluate our framework along three axes: the accuracy of learned bias probe, multi-group fairness estimation via bias probes and the sample and computation efficiency of the proposed interactive procedure. Given a dataset D, we first train three classifiers h —linear model, random forest, or multi-layer perceptron (MLP). We then generate a labeled dataset by querying h on unlabeled samples. Using this dataset, we perform interactive learning of bias probes associated with multi-group fairness. From the learned probes, we estimate the fairness metric $\mu(\mathcal{D}, h)$ and report the estimation error along with explanations for the audit report.

Datasets and experimental details. We evaluate our theoretical results on three well-known benchmark datasets: COMPAS, GERMAN CREDIT and STUDENT. For the COMPAS dataset, introduced in the context of recidivism risk assessment Angwin et al. (2022), the number of groups is fixed to $k = 2$, while for the GERMAN CREDIT dataset, originating from the Statlog German Credit benchmark dataset Hofmann (1994)., we consider $k = 4$ groups. In Appendix H, we further extend our analysis to synthetic datasets with multiple values of k.

Evaluation details and reproducibility. As a benchmark, we compare k-ALEBI against the two baselines, DIRECTAUDIT and RECONAUDIT. For each binary classifier, we construct a dataset and use it as a pool of unlabeled samples, which we partition into k group-specific pools. k-ALEBI uses these pools to selectively construct informative CGQ queries. Each experiment is run with 10 random seeds to ensure reproducibility.

4.1 ACCURACY

Across the three datasets, k-ALEBI is competitive with the baselines of DIRECTAUDIT and RECONAUDIT in the accuracy of multi-group fairness estimation while requiring only few CGQs. As Figure 1 shows, after 1000 queries k-ALEBI generally achieved substantially smaller estimation error than the the baselines, although reconstruction could achieve essentially zero error when it successfully recovered the target within the finite candidate class.

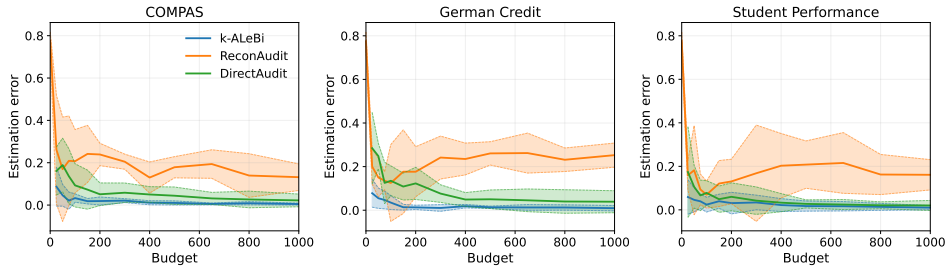

Figure 1: Estimation Error of Multi-group fairness under limitted budget.

We further evaluate k-ALEBI against DIAMAUDIT, the auditing method proposed by Yan & Zhang (2022), in its applicable setting of $k = 2$ protected groups and a linear hypothesis class, corresponding to statistical parity. Figure 2 shows that k-ALEBI achieves lower statistical parity estimation error than DIAMAUDIT on both datasets. On COMPAS, k-ALEBI achieves an estimation error of 0.0000213, compared with 0.000046 for DIAMAUDIT. On the Student dataset, k-ALEBI achieves an estimation error of 0.006947.

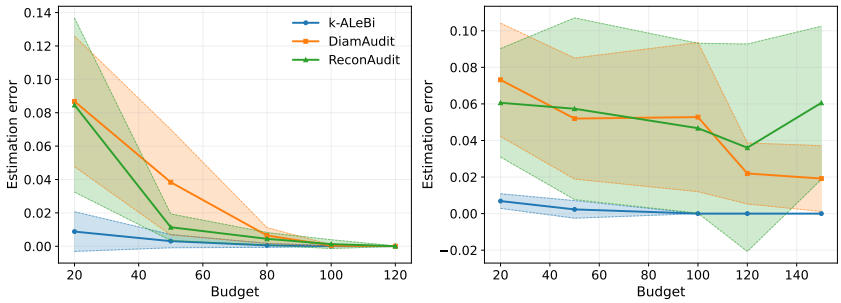

Figure 2: Estimation Error of DIAMAUDIT under Its Applicable Setting

Due to space constraints, we defer additional experimental results for interpretability, model extraction attacks, and robustness against fairness-aware adversary to Appendix H.

5 CONCLUSION

In this work, we established an explicit separation between reconstruction-based and probe-based approaches to fairness auditing. We characterized the probe-learning problem through the multi-context star number and showed that, even when reconstruction-based auditing is infeasible for linear classifiers, the corresponding probe-based approach remains learnable, addressing **Q1**. Building on this formulation, we designed an auditing algorithm based on learning interpretable bias probes, providing audit information directly in terms of the fairness property of interest and addressing **Q2**. Finally, we analyzed the auditing problem in the presence of a fairness-aware adversary and established both upper and lower bounds on the resulting audit complexity, addressing **Q3**.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Nir Ailon. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13(1), 2012.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 65–72, 2006.

Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active property testing. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 21–30. IEEE, 2012.

Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pp. 642–653, 2020.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's -contamination model. 2016.

Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting. *Advances in Neural Information Processing Systems*, 36:6070–6091, 2023.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31): eabq6147, 2022.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 20, 2007.

Martin Ebers. Truly risk-based regulation of artificial intelligence how to implement the eu's ai act. *European Journal of Risk Regulation*, 16(2):684–703, 2025.

European Commission. The digital services act package. Technical report, European Union, 2024a. URL https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

European Commission. Regulation on artificial intelligence (ai act), 2024b. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Final text adopted by the European Parliament.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Ana Cristina Bicharra Garcia, Marcio Gomes Pinto Garcia, and Roberto Rigobon. Algorithmic discrimination in the credit domain: what do we know about it? *AI & society*, 39(4):2059–2098, 2024.

Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Drew Harwell. A face-scanning algorithm increasingly decides whether you deserve the job. In *Ethics of Data and Analytics*, pp. 206–211. Auerbach Publications, 2022.

Hans Hofmann. Statlog (german credit data). *UCI Machine Learning Repository*, 10:C5NC77, 1994.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1839–1851, 2022.

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

Daniel M Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 355–366. IEEE, 2017.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Allen Liu and Ankur Moitra. Model stealing for any low-rank language model. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1755–1761, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14653–14671, 2024.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. h-consistency bounds for pairwise misranking loss surrogates. In *International conference on Machine learning*, pp. 23743–23802. PMLR, 2023.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pp. 3452–3482. PMLR, 2021.

Kirsten N Morehouse, Siddharth Swaroop, and Weiwei Pan. Rethinking llm bias probing using lessons from the social sciences. *arXiv preprint arXiv:2503.00093*, 2025.

New York State Senate. S8612: An act to regulate automated decision-making tools, 2024. URL https://www.nysenate.gov/legislation/bills/2023/S8612.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4954–4963, 2019a.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against dnn model stealing attacks. *arXiv preprint arXiv:1906.10908*, 2019b.

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pp. 1907–1916. PMLR, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Joel Tropp. Freedman's inequality for matrix martingales. 2011.

Keyon Vafa, Peter G Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. *arXiv preprint arXiv:2507.06952*, 2025.

Amey Vrudhula, Alan C Kwan, David Ouyang, and Susan Cheng. Machine learning and bias in medical imaging: opportunities and challenges. *Circulation: Cardiovascular Imaging*, 17(2): e015495, 2024.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1578–1590, 2024.

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

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227. IEEE, 1977.

Xiaojian Yuan, Kejiang Chen, Wen Huang, Jie Zhang, Weiming Zhang, and Nenghai Yu. Data-free hard-label robustness stealing attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6853–6861, 2024.

Xiaoyong Yuan, Leah Ding, Lan Zhang, Xiaolin Li, and Dapeng Oliver Wu. Es attack: Model stealing against deep neural networks without data hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1258–1270, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

A Additional Context and Related Works

In addition to the limitation discussed in the introduction, combined with increasing regulatory pressure, have shifted attention toward post-deployment auditing. In the United States, regulations such as New York City's Local Law 144 New York State Senate (2024) mandate bias audits of automated decision systems. In Europe, the Digital Services Act European Commission (2024a) and the AI Act European Commission (2024b); Ebers (2025) impose transparency, risk management, and auditing requirements for large-scale and high-risk ML systems in online platforms. These developments reflect a broader shift, where auditing is becoming a central requirement for deploying ML systems in high-stakes settings. In this post-deployment regime, the model is typically treated as a black box. An external auditor seeks to assess the model's behavior without access to its internal structure, interacting with the model owner through queries. This setting introduces a fundamental tension: the auditor aims to extract sufficient information to reliably evaluate fairness, while the model owner may wish to limit disclosure in order to protect model's sensitive information.

The following figure summarizes our setting and the objectives we pursue.

In Figure 3, the auditor is given the task to audit a property of interest μ, has access to a pool of unlabeled instances $\mathcal{Q}$ sampled from the marginal input distribution, providing information about the regions of the input space relevant to the audit. The auditor then selectively queries the model owner with samples that are informative with respect to μ. At round $t+1$, the auditor constructs a query q_{t+1} and sends it to the model owner $\mathbf{M}$, who returns a response $\mathbf{M}_{t+1} = \mathbf{M}(q_{t+1})$.

We assume a powerful, fairness-aware adversary that may strategically alter its responses to conceal disparities across groups. Based on the history $\{(q_1, \mathbf{M}_1), \ldots, (q_{t+1}, \mathbf{M}_{t+1})\}$, the auditor produces

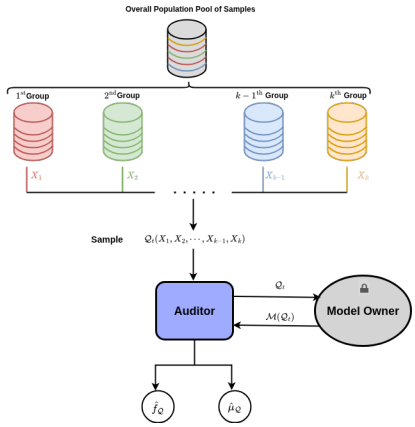

Figure 3: Interactive fairness auditing against a fairness-aware adversary.

an estimate $\hat{\mu}_{t+1}$ of the property under audit, together with regions of the input space in which the model exhibits larger fairness disparities.

In this work, we address these challenges in the setting of multi-group fairness. We consider an interactive auditing framework in which an auditor queries a black-box model through carefully designed inputs drawn from a pool of unlabeled data. Our approach focuses on learning comparison functionals that capture the model's discriminative behavior across protected groups, rather than reconstructing the model itself. This perspective enables both statistical efficiency and interpretability while inherently limiting the information exposed to the auditor, thereby protecting the model owner from reconstruction attacks.

In Figure 3, the auditor is given the task to audit a property of interest μ, has access to a pool of unlabeled instances $\mathcal{Q}$ sampled from the marginal input distribution, providing information about the regions of the input space relevant to the audit. The auditor then selectively queries the model owner with samples that are informative with respect to μ. At round $t+1$, the auditor constructs a query q_{t+1} and sends it to the model owner $\mathbf{M}$, who returns a response $\mathbf{M}_{t+1} = \mathbf{M}(q_{t+1})$.

A.1 Additional Related Work

Active Learning and Disagreement-Based Complexity. Disagreement-based active learning characterizes label complexity through geometric and combinatorial quantities such as the disagreement coefficient in distribution-dependent settings (Balcan et al., 2006; Dasgupta et al., 2007) and the star number in distribution-free regimes (Hanneke, 2014; Hanneke & Yang, 2015). These parameters capture the ability of a hypothesis class to isolate individual instances via localized perturbations around a reference hypothesis. Our analysis is inspired by this framework but departs from it in a fundamental way. In classical active learning, disagreement regions are subsets of the instance space. In contrast, the auditing setting we consider induces disagreement over cross-group tuples, leading to a multipartite geometry rather than a pointwise one. To capture this structure, we introduce the intergroup star number, which extends the classical notion of star number to relational settings between sub-populations. This quantity measures the capacity of a comparison class to isolate cross-group configurations, and serves as the key complexity parameter governing sample efficiency in our setting.

Preference Learning. Preference learning is closely related to our setting, as it also relies on pairwise comparisons. In its classical formulation, a collection of users provides pairwise preferences over a finite set of items, and the goal is to predict unseen preferences. Many approaches aim to estimate a score matrix of size (number of users) $\times$ (number of items), often under low-rank assumptions and solved via convex optimization Park et al. (2015). Several works consider active querying in this setting. For instance, Ailon (2012) studies adaptive pairwise comparisons over a finite set of items and returns a total order whose disagreement is competitive with the optimal ranking. However, their guarantees concern approximation to the best ranking with bounded query

complexity, rather than uniform learning over a general function class. Similarly, methods based on convex relaxations, such as SVM-rank, operate within a fixed parametric class and focus on ranking performance rather than learning comparison functionals. Mao et al. (2023) provides finite-sample guarantees linking surrogate losses to target ranking losses. However, this line of work focuses on bipartite ranking and misranking loss, rather than recovering structure from pairwise comparisons in a general interactive setting. In contrast, our approach does not aim to learn a ranking over a finite set of items. Instead, we learn a class of comparison functionals that capture the discriminative behavior of a black-box model over the input distribution. This shifts the problem from ranking to distributional estimation, where the goal is to approximate the induced comparison structure with uniform guarantees over a function class.

Model Extraction Attacks. Model extraction attacks have emerged as a major threat to ML systems deployed through black-box interfaces. In these attacks, an adversary interacts with a model via query access and trains a surrogate model that replicates the functionality of the original model (Tramèr et al., 2016; Orekondy et al., 2019a; Yuan et al., 2024; Orekondy et al., 2019b). Early work demonstrated that even simple query strategies can recover models such as logistic regression, decision trees, and neural networks with high fidelity (Tramèr et al., 2016). Subsequent work has shown that extraction is possible even without access to real data, by generating synthetic queries or by using knowledge distillation techniques (Orekondy et al., 2019a; Yuan et al., 2022). From a broader perspective, model extraction raises both intellectual property and security concerns, as stolen models can be reused or further exploited for downstream attacks (Jagielski et al., 2020). In the context of large language models (LLMs), model owners with API access are highly vulnerable to model extraction attacks, posing significant security risks. In particular, Liu & Moitra (2025) show that any low-rank language model can be learned in polynomial time using conditional queries. We design auditing mechanisms that intentionally restrict the information revealed through queries, and show that recovering the underlying model from our bias probes is computationally hard (under standard assumptions).

A.2 Interpretable Audit Report

We consider the following classical example in fairness, which illustrates the value of interpretability in providing clear and compelling evidence of a model's harmful or discriminatory behavior.

Example 1 (Interpretable Audit Report via Unprotected Feature-wise Bias Probes)**.** *Consider a credit-lending setting in which a bank deploys a binary classifier $h : \mathcal{X} \to \{0, 1\}$ to decide whether to grant a loan. Suppose that each client is represented by three unprotected input features, with monthly income and age corresponding to the first and second coordinates, respectively, and let gender be the protected attribute. For simplicity, assume two protected groups. A cross-group query is $Q = (X_1, X_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, where X_1 and X_2 are sampled from the two protected groups. The associated relational bias probe is $f_{1,2}(Q) = h(X_1) - h(X_2) \in \{-1, 0, +1\}$. The auditor observes this relational response rather than the two individual predictions separately.*

For an unprotected feature q, let $P_q(Q) := \left(X_1^{(q)}, X_2^{(q)}\right)$. For example, $P_1(Q) = \left(X_1^{(1)}, X_2^{(1)}\right)$ contains the monthly incomes of the two queried clients, while $P_2(Q) = \left(X_1^{(2)}, X_2^{(2)}\right)$ contains their ages. The feature-wise bias profile is $\mathfrak{e}_q(z_1, z_2) := \mathbb{E}[f_{1,2}(Q) \mid P_q(Q) = (z_1, z_2)]$, and its unsigned magnitude is $\widetilde{f}_q(z_1, z_2) := |\mathfrak{e}_q(z_1, z_2)|$. Thus, $\mathfrak{e}_q$ records the direction and magnitude of the conditional group prediction disparity, while $\widetilde{f}_q$ records only its magnitude.

The resulting audit report can combine the following complementary summaries:

- ***Global bias probe.*** *The global statistical-parity disparity is obtained from $\mu(h) = |\mathbb{E}[f_{1,2}(Q)]|$. This provides an overall measure of the prediction-rate difference between the two protected groups.*

- ***Feature-wise bias profile.*** *For a feature such as age or monthly income, the function m_q describes how the signed relational disparity varies across values of that feature. For example, $\mathfrak{e}_1(z_1, z_2)$ gives the expected relational response among cross-group queries whose two clients have monthly incomes near (z_1, z_2).*

- ***Expected feature-wise bias magnitude.*** *The quantity* $\widetilde{f}_q(z_1, z_2) = |\mathfrak{e}_q(z_1, z_2)|$ *identifies regions of the projected feature space in which the conditional group disparity is large. For monthly income, it can therefore reveal income combinations for which the model exhibits particularly pronounced group-relative prediction differences; the same construction applies to age.*

Importantly, these quantities provide probe-level interpretability*: they are estimable from projected CGQ feature values and relational probe responses. They should not be interpreted as ordinary predictive feature importance or as causal effects of the corresponding features on the classifier.*

Bias probes thus provide a localized view of the model's discriminative behavior, isolating how differences across groups manifest at specific feature configurations. While, the expected feature-wise bias probe represents the average intergroup comparison induced by h among all inputs sharing the same configuration on the unprotected feature q, providing a low-dimensional summary of the model's discriminative behavior.

B The Fairness Auditing Landscape: Existing Approaches

B.1 Audit via Model Reconstruction

In this setting, the auditor has access to the hypothesis class $\mathcal{H}$ and query access to the model h, i.e., $\mathrm{I}_{\mathbf{M}}(h) = \{\mathcal{H}, \mathbf{M}(h)\}$. The goal is to reconstruct a surrogate model by emulating the training process within $\mathcal{H}$. For instance, Yan & Zhang (2022) use disagreement queries to identify a subclass of $\mathcal{H}$ whose elements share the same statistical parity on a given unlabeled pool.

Interaction Protocol for Reconstruction-based Audit.

The interaction between the auditor ReconAudit and the model owner $\mathbf{M}(h)$ proceeds as follows:

- The auditor is given the hypothesis class $\mathcal{H}$ and access to the unlabeled pool $\mathcal{Q}$, and a query budget T.
- **Round** $t = 1$**.** The auditor selects a query $X_1 \in \mathcal{Q} \cap \mathrm{DIS}(\mathcal{H})$ and sends it to the model owner. The model owner returns the response $h(X_1)$.
- **Round** $t = 2, \cdots, T$**.** Based on the history of previous queries and responses, the auditor updates its current hypothesis class V_t and selects a query $X_t \in \mathcal{P} \cap \mathrm{DIS}(V_t)$. It sends X_t to the model owner and receives the response $h(X_t)$.
- **Reconstruction.** After T rounds, the auditor selects any $\hat{h} \in V_T$ to reconstruct a surrogate model of h.
- **Audit.** The auditor applies a plug-in estimator to $\hat{h}$ to estimate the multi-group fairness property using the history of queries collected during the interaction.

However, such approaches rely on queries designed for learning rather than for estimating the target property. As a result, their sample complexity scales with that of learning, typically $\mathcal{O}(d/\varepsilon^2 \log \frac{1}{\delta})$, where d denotes a complexity measure of $\mathcal{H}$ (e.g., VC dimension). This dependence not only leads to inefficiency but also raises model confidentiality concerns, as it may enable model extraction. We design our first baseline, ReconAudit, inspired by the reconstruction-based auditing approach.[2]

[2]We note that the algorithm of (Yan & Zhang, 2022) is restricted to the class of linear separators. We therefore develop a general reconstruction-based algorithm that does not rely on a specific model class.

Algorithm 2 RECONAUDIT

Require: $\mathrm{I}_{\mathbf{M}}(h) = \{\mathcal{H}, \mathbf{M}(h)\}$, query pool $\mathcal{Q}$, budget T

1: Initialize
$$V_0 = \mathcal{H}.$$
2: **for** $t = 1, \ldots, T$ **do**
3: Select
$$X_t \in \mathrm{DIS}(V_t)$$
4: Query
$$\mathbf{M}_t = h(X_t).$$
5: Update
$$V_t = \{h' \in V_{t-1} : h'(X_t) = \mathbf{M}_t\}.$$
6: **end for**
7: Reconstruct
$$\widehat{h} \in V_T$$
8: **return** $\widehat{h}$ and
$$\widehat{\mu}_T = \mu_{\mathcal{Q}_{1:T}}(\widehat{h}).$$

Remark 8. *We first observe that model learning does not depend on the protected groups represented in the data. In particular,* RECONAUDIT *does not account for group membership during the learning procedure, which may lead to higher reconstruction errors for some groups than for others.*

Second, once the model has been reconstructed, the auditor effectively has white-box access to the model, and the subsequent audit can therefore employ any white-box procedure, including procedures for producing interpretable audit reports. Whether such a reconstruction-based strategy remains robust in the presence of a fairness-aware adversary, however, is not clear a priori. *We investigate this question in Section 4.*

B.2 AUDIT VIA DIRECT ESTIMATION

In this setting, the auditor samples $\mathcal{O}(1/\varepsilon^2 \log \frac{1}{\mu})$ from each group in the pool $\mathcal{P}$, and perform membership queries [3]. The objective is to estimate the target property μ directly from these samples, without reconstructing the model. While the sample complexity still scales as $\mathcal{O}(k/\varepsilon^2 \log \frac{1}{\mu})$, it no longer depends on the complexity of $\mathcal{H}$. This avoids the need for uniform approximation over the hypothesis class and mitigates risks associated with model reconstruction.

Interaction Protocol for Direct-Estimation Audit

The interaction between the auditor DIRECTAUDIT and the model owner $\mathbf{M}(h)$ proceeds as follows:

- The auditor is given access to an unlabeled pool $\mathcal{Q}$ and a query budget T.
- **Rounds** $t = 1, \ldots, T$**.** At each round, for each group $i \in [k]$, the auditor samples points from $\mathcal{Q}$ and queries the model owner. It uses the responses to estimate the expected positive-label rate for each group.
- **Estimation.** After T rounds, the auditor obtains estimates of the pairwise disparities
$$\{\hat{\mu}_{ij} : i, j \in [k],\, i < j\},$$
forming a vector of size $\frac{k(k-1)}{2} = \mathcal{O}(k^2)$.
- **Audit.** The auditor outputs the maximum estimated pairwise disparity as its estimate of the multi-group fairness property.

Remark 9. *The reconstruction-based audit can be viewed as a* search-then-estimate *procedure: the auditor first searches for a surrogate model consistent with its observations and only then extracts*

[3] or alternatively has access to a random query oracle

the audit report from the reconstructed model. In contrast, the direct-estimation approach estimates the fairness property directly from queried responses, without first reconstructing the model.

This distinction is particularly relevant when the auditor is used to verify whether a model update changes the property under audit. In the reconstruction-based approach, the learned surrogate can be used to check whether the updated model agrees with the pre-audit model on the set of points used for the plug-in estimation. However, this set is also used during model reconstruction and has size scaling as $\mathcal{O}\left(\frac{d}{\varepsilon^2}\log\frac{1}{\mu}\right)$*. Consequently, requiring the updated model to agree with the original model on this large set of points can impose a substantially stronger constraint than requiring the fairness property itself to remain unchanged. In this sense, the reconstruction-based audit may effectively constrain the model update over a large portion of the input space, rather than merely verifying the stability of the property under audit.*

Algorithm 3 DIRECTAUDIT

Require: $\mathrm{I}_{\mathbf{M}}(h) = \mathbf{M}(h)$, Query distribution $\mathcal{D}$, budget T

1: **for** $i = 1, \ldots, k$ **do**
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $X_t^{(i)} \sim \mathcal{D}_{\cdot|\mathcal{Z}_i}$ ▷ $\mathcal{D}_{\cdot|\mathcal{Z}_i}$: marginal distribution conditioned on group i.
4: Membership Query $\mathbf{M}_t^{(i)} = h(X_t^{(i)})$ ▷ Alternatively, use random queries from the joint distribution.
5: **end for**
6: Estimate model behavior on every group

$$\widehat{\mu}_T^{(i)} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{M}_t^{(i)}.$$

7: **end for**
8: Compute every pairwise component

$$\mu = \left(\widehat{\mu}_T^{(i)} - \widehat{\mu}_T^{(j)}\right)_{1\le i<j\le k}$$

9: Return

$$\widehat{\mu} = ||\mu||_\infty$$

B.3 MUTLI-STAR NUMBER OF MULTI-CONTEXT CLASS RELATIVE TO LINEAR CLASSIFIERS

Proposition 2 (Information-Theoretic Separation of Reconstruction and Probe Auditing)**.** *There exists an audit instance problem, where auditing via model reconstruction is strictly harder than auditing via probe learning.*

Proof

Let $\mathcal{H} = \left\{h_w(X) = \mathbb{1}_{wX\ge 0} : w \in \mathbb{R}^{*2}\right\}$ be the class of homogeneous linear classifiers.

Let $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ denote rotation by $\pi/2$, and consider the following representation of protected groups domain $\mathcal{X}_R = \mathcal{Z}_1 \times \mathcal{Z}_2 = \{(x, Rx) : x \in \mathbb{R}^{*2}\}$.

For $h \in \mathcal{H}$, we have the bias probe for our audit instance problem $k = 2$, $f_h(x, Rx) = h(x) - h(Rx) \in \{-1, 0, +1\}$, and let $\mathcal{F}_{R,\mathcal{H}} = \{f_h : h \in \mathcal{H}\}$.

We show that $\mathfrak{s}(\mathcal{H}) = \infty, \qquad \mathfrak{s}_2(\mathrm{F}_{\mathcal{H},R}) = 3.$

The fact that $\mathfrak{s}(\mathcal{H}) = \infty$ for homogeneous linear separators in $\mathbb{R}^2$ is standard (Hanneke, 2014). Thus it remains to compute the star number of the induced relational class $\mathcal{F}_{\mathcal{H},R}$.

Step 1: Angular representation of the 2-context class. Since the classifiers are homogeneous, only the directions of w and x matter.

Let $w_\theta = (\cos\theta, \sin\theta)$ and $x_\phi = r(\cos\phi, \sin\phi)$, where $r > 0$.

Then $w_\theta x_\phi = r\cos(\theta - \phi)$. Since Rx_ϕ has angle $\phi + \pi/2$,

$$w_\theta^\top R x_\phi = r\cos\left(\theta - \phi - \frac{\pi}{2}\right) = r\sin(\theta - \phi).$$

Consequently,

$$f_\theta(\phi) = \operatorname{sign}\big(\cos(\theta - \phi)\big) - \operatorname{sign}\big(\sin(\theta - \phi)\big). \tag{1}$$

We choose all probe and witness angles away from the transition points, so the value assigned to angle 0 is irrelevant.

Let $\beta = \theta - \phi \pmod{2\pi}$. Then,

$$f_\theta(\phi) = \begin{cases} +1, & \beta \in (-\pi/2, 0), \\ 0, & \beta \in (0, \pi/2), \\ -1, & \beta \in (\pi/2, \pi), \\ 0, & \delta \in (\pi, 3\pi/2). \end{cases} \tag{2}$$

Thus the 2-context class becomes the β translation class obtained by rotating the four-sector cyclic labeling $+1, \quad 0, \quad -1, \quad 0$, with each sector having angular width $\pi/2$.

Step 2: Agreement sets relative to a probe. Fix a reference orientation θ_0. By rotational invariance, we may assume without loss of generality that $\theta_0 = 0$. For every probe angle ϕ, define its agreement set in parameter space by

$$A_\phi = \left\{\theta \in \mathbb{S}^1 : f_\theta(\phi) = f_0(\phi)\right\}.$$

By construction, $0 \in A_\phi$ for every ϕ. Equation equation 2 shows that there are exactly two possible forms for A_ϕ.

- *Discriminatory Probe:* $f_0(\phi) \in \{-1, +1\}$*:* In this case the reference label occurs on exactly one sector of length $\pi/2$. Hence $A_\phi = I_\phi$, where I_ϕ is an arc of length $\pi/2$ containing 0.
- *Fair Probe:* $f_0(\phi) = 0$. The label 0 occurs on two antipodal sectors, each of length $\pi/2$. Therefore $A_\phi = I_\phi \cup (I_\phi + \pi)$, where again I_ϕ is the unique component containing 0 and has length $\pi/2$.

Thus every probe has a *local agreement interval* I_ϕ of length $\pi/2$ containing the reference parameter 0. A Fair probe additionally has an antipodal copy of this interval on the two dimensional sphere.

Step 3: An interval lemma. We use the following elementary fact.

Lemma 10. *Let*

$$I_i = (\ell_i, r_i), \qquad i = 1, \ldots, m,$$

be intervals on the real line satisfying $\ell_i < 0 < r_i$ for every i. Suppose that, for an index i, there exists a point t_i such that

$$t_i \in \bigcap_{j \neq i} I_j \qquad \textit{but} \qquad t_i \notin I_i.$$

Then i must be either an index attaining the largest left endpoint or an index attaining the smallest right endpoint. Consequently, at most two indices can satisfy this property.

Proof Suppose $t_i \in \bigcap_{j \neq i} I_j \setminus I_i$. Since every interval contains 0, either $t_i < \ell_i < 0$ or $t_i > r_i > 0$.

If $t_i < \ell_i$, then $t_i \in I_j$ for every $j \neq i$, so $\ell_j < t_i < \ell_i$ for all $j \neq i$. Hence $\ell_i > \ell_j$ for all $j \neq i$, so i is the unique index with largest left endpoint.

Similarly, if $t_i > r_i$, then $r_i < t_i < r_j$ for all $j \neq i$, and therefore i is the unique index with smallest right endpoint.

There can be at most one unique largest left endpoint and at most one unique smallest right endpoint. Hence at most two indices can possess such a private point. ■

Step 4: Upper bound $\mathfrak{s}(\mathcal{F}air) \leq 3$. Suppose that $\phi_1, \ldots, \phi_m$ form a star centered at f_0. By the definition of the star number, for every $i \in [m]$ there exists an orientation θ_i such that

$$f_{\theta_i}(\phi_i) \neq f_0(\phi_i), \qquad f_{\theta_i}(\phi_j) = f_0(\phi_j) \quad \text{for all } j \neq i. \tag{3}$$

Equivalently,

$$\theta_i \in \bigcap_{j \neq i} A_{\phi_j} \setminus A_{\phi_i}. \tag{4}$$

We distinguish cases according to the number of discriminatory probes.

- *Case 1: There are no discriminatory probes.*
  Then every probe is fair:
  $$A_{\phi_i} = I_i \cup (I_i + \pi).$$
  Since this family is invariant under translation by π, we may project the parameter circle modulo π. Under this projection, each agreement set becomes a single interval I_i of length $\pi/2$ containing the reference point 0.
  Condition equation 4 therefore reduces exactly to the private-point condition in Lemma 10. Consequently, $m \leq 2$.
- *Case 2: There is exactly one discriminatory probe.*
  Let this probe have index q. For every $i \neq q$, a witness θ_i satisfying equation 4 must in particular belong to $A_{\phi_q} = I_q$. Since I_q is the local interval of length $\pi/2$ containing 0, θ_i cannot lie in the antipodal component $I_j + \pi$ of any Fair agreement set. Hence, for all $i \neq q$, the star condition reduces locally to
  $$\theta_i \in \bigcap_{j \neq i} I_j \setminus I_i.$$
  By Lemma 10, at most two of the Fair probes can possess such witnesses.
  Since there is only one discriminatory probe, this yields $m \leq 1 + 2 = 3$.
- *Case 3: There are at least two Discriminatory probes.*
  For every i, the intersection
  $$\bigcap_{j \neq i} A_{\phi_j}$$
  contains a Discriminatory agreement interval, except possibly when i is one of the Discriminatory probes and there is exactly one other Discriminatory probe. In either case, because there is at least one Discriminatory agreement set among the sets indexed by $j \neq i$, every witness θ_i is confined to the local component containing 0.
  Thus all relevant agreement sets may be replaced by their local intervals I_j, and condition equation 4 again reduces to
  $$\theta_i \in \bigcap_{j \neq i} I_j \setminus I_i.$$
  Lemma 10 then implies $m \leq 2$.

Combining the three cases gives $\mathfrak{s}_2(\mathcal{F}_{\mathcal{H},R}) \leq 3$.

Step 5: Lower bound $\mathfrak{s}_2(\mathcal{F}_{\mathcal{H},R}) \geq 3$. It remains to exhibit an explicit 3-star on the 2-context class. To this end, consider the reference orientation $\theta_0 = 0$ and choose the three probe angles

$$\phi_1 = \frac{\pi}{20}, \qquad \phi_2 = \frac{21\pi}{40}, \qquad \phi_3 = \frac{23\pi}{40}.$$

Using equation 1, their labels under the reference orientation are

$$(f_0(\phi_1), f_0(\phi_2), f_0(\phi_3)) = (+1, 0, 0).$$

Now choose the three witnessing orientations

$$\theta_1 = \frac{4\pi}{5}, \qquad \theta_2 = \frac{3\pi}{80}, \qquad \theta_3 = \frac{25\pi}{16}.$$

Direct substitution into equation 1 gives

$$(f_{\theta_1}(\phi_1), f_{\theta_1}(\phi_2), f_{\theta_1}(\phi_3)) = (-1, 0, 0),$$
$$(f_{\theta_2}(\phi_1), f_{\theta_2}(\phi_2), f_{\theta_2}(\phi_3)) = (+1, +1, 0),$$

and

$$(f_{\theta_3}(\phi_1), f_{\theta_3}(\phi_2), f_{\theta_3}(\phi_3)) = (+1, 0, -1).$$

Therefore,

$$\{j : f_{\theta_1}(\phi_j) \neq f_0(\phi_j)\} = \{1\},$$
$$\{j : f_{\theta_2}(\phi_j) \neq f_0(\phi_j)\} = \{2\}$$

and

$$\{j : f_{\theta_3}(\phi_j) \neq f_0(\phi_j)\} = \{3\}$$

Hence $\mathfrak{s}(\mathcal{F}_{\mathcal{H},R}) \geq 3$. Together with the upper bound, we get $\mathfrak{s}_2(\mathcal{F}_{\mathcal{H},R}) = 3$.

Finally, since $\mathfrak{s}(\mathcal{H}) = \infty$, we obtain the desired information-theoreticseparation. ■

Remark 11 (The protected groups domain restriction is essential)**.** *The finite-star conclusion on the 2-context class relies essentially on restricting the allowable cross-group probes to*

$$\mathcal{Z}_R = \{(x, Rx) : x \neq 0\}.$$

It does not hold for the unrestricted product-space class

$$\widetilde{\mathcal{F}} = \{(x_1, x_2) \mapsto h(x_1) - h(x_2) : h \in \mathcal{H}\}.$$

Indeed, for every homogeneous linear classifier and every x away from its decision boundary,

$$h(-x) = -h(x),$$

and hence

$$f_h(x, -x) = h(x) - h(-x) = h(x)^4.$$

Thus the restriction of $\widetilde{\mathcal{F}}$ to pairs of the form $(x, -x)$ contains an exact copy of $\mathcal{H}$, implying

$$\mathfrak{s}_2(\widetilde{\mathcal{F}}) \geq \mathfrak{s}(\mathcal{H}) = \infty.$$

Therefore the separation in Proposition 2 arises from the combination of relational predictions and the structured protected groups domain, rather than from taking prediction differences alone.

C TECHNICAL LEMMAS

Proposition 1 (Coupling Invariance)**.** *Let μ denotes multi-group fairness defined in 1.*

Let $\nu \in \Pi(\nu_1, \ldots, \nu_k)$ be any coupling measure on $\mathcal{X}$ whose ith marginal satisfies $\nu_i(\cdot) = \mathcal{D}(\cdot \mid \mathcal{Z}_i)$, for all $i \in [k]$. The following holds:

$$\mu(h, \mathcal{D}) = \|\mathbb{E}_{\mathbf{X}\sim\nu}[f_h(\mathbf{X})]\|_\infty$$

Proof Let $\nu \in \Pi(\nu_1, \ldots, \nu_k)$ be any coupling measure on $\mathcal{X}$ whose ith marginal satisfies $\nu_i(\cdot) = \mathcal{D}(\cdot \mid \mathcal{Z}_i)$, for all $i \in [k]$ For $i \neq j$, let ν_{ij} denote the (i, j) marginal of ν on $\mathcal{Z}_i \times \mathcal{Z}_j$.

By definition, $f_h(\mathbf{X}) = \big(h(X_i) - h(X_j)\big)_{1\leq i<j\leq k}$.

[4]Here, we reorder the labels: instead of $\{-2, 0, 2\}$ (h takes labels $\{0, 1\}$) we use labels $\{-1, 0, 1\}$ (h takes labels $\{-1, 1\}$))

Since expectation of a finite-dimensional random vector is taken componentwise,

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim\nu}\left[f_h(\mathbf{X})\right] &= \left(\mathbb{E}_{\mathbf{X}\sim\nu}\left[h(X_i)-h(X_j)\right]\right)_{1\le i<j\le k}\\ &= \left(\mathbb{E}_{(X_i,X_j)\sim\nu_{ij}}\left[h(X_i)-h(X_j)\right]\right)_{1\le i<j\le k}\\ &= \left(\int_{\mathcal{Z}_i\times\mathcal{Z}_j}\big(h(x_i)-h(x_j)\big)\,d\nu_{ij}(x_i,x_j)\right)_{1\le i<j\le k}\\ &= \left(\int_{\mathcal{Z}_i\times\mathcal{Z}_j}h(x_i)\,d\nu_{ij}(x_i,x_j)-\int_{\mathcal{Z}_i\times\mathcal{Z}_j}h(x_j)\,d\nu_{ij}(x_i,x_j)\right)_{1\le i<j\le k}.\end{aligned}$$

Because the marginals of ν_{ij} are ν_i and ν_j, respectively,

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim\nu}\left[f_h(\mathbf{X})\right] &= \left(\int_{\mathcal{Z}_i}h(x)\,d\nu_i(x)-\int_{\mathcal{Z}_j}h(x)\,d\nu_j(x)\right)_{1\le i<j\le k}\\ &= \left(\mathbb{E}_{X\sim\mathcal{D}(\cdot|\mathcal{Z}_i)}[h(X)]-\mathbb{E}_{X\sim\mathcal{D}(\cdot|\mathcal{Z}_j)}[h(X)]\right)_{1\le i<j\le k}.\end{aligned}$$

And because $\mu_{ij}(h,\mathcal{D}) := \mathbb{E}_{X\sim\mathcal{D}(\cdot|\mathcal{Z}_i)}[h(X)]-\mathbb{E}_{X\sim\mathcal{D}(\cdot|\mathcal{Z}_j)}[h(X)]$, we therefore obtain

$$\mathbb{E}_{\mathbf{X}\sim\nu}[f_h(\mathbf{X})] = \big(\mu_{ij}(h,\mathcal{D})\big)_{1\le i<j\le k}$$

Consequently, $\|\mathbb{E}_{\mathbf{X}\sim\nu}[f_h(\mathbf{X})]\|_\infty = \max_{1\le i<j\le k}|\mu_{ij}(h,\mathcal{D})|$

Therefore

$$\mu(h,\mathcal{D}) = \|\mathbb{E}_{\mathbf{X}\sim\nu}[f_h(\mathbf{X})]\|_\infty$$

■

This reformulation shows that auditing multi-group fairness can be viewed as estimating expectations of pairwise comparison functionals under the product distribution. In particular, it shifts the objective from learning the model h itself to estimating the structured family of functional defined on $\mathcal{X}$. The resulting interaction model differs fundamentally from reconstruction-based approaches: rather than approximating h over the hypothesis class $\mathcal{H}$, the auditor directly probes the relational structure induced by h across protected groups.

Remark 12. *Although the value of the audited property is invariant to the choice of coupling, the coupling determines the distribution of the relational queries observed by the auditor and therefore the information available during the audit. We therefore treat the coupling as part of the audit design: a coupling $\nu \in \Pi(\mathcal{D}_1,\dots,\mathcal{D}_k)$ is selected before interaction and remains fixed throughout the auditing procedure, with CGQs sampled i.i.d. from ν. The auditor may adaptively decide which sampled CGQs to query, but may not adapt the coupling itself in response to previous oracle outputs. This restriction does not constrain the value of the fairness functional, which depends only on the protected-group marginals, but it fixes the statistical experiment through which that functional is learned. Allowing the coupling to change adaptively would define a strictly stronger query model in which the auditor can modify the distribution of future CGQs based on past responses, rather than merely perform active selection from a fixed data-generating distribution.*

C.1 Information Complexities

For a fixed multi-context class $\mathcal{F}$, the disagreement set is defined as follows: $\mathrm{DIS}(\mathcal{F}) \triangleq \{\mathbf{x}\in\mathcal{X} : \exists f,f'\in\mathcal{F} : f(\mathbf{x})\neq f'(\mathbf{x})\}$. The disagreement sets arising in learning and auditing are objects of a different nature. In standard learning, the disagreement region is a subset of the sample space $\mathcal{Z}$, reflecting uncertainty about the label of individual instances. In contrast, for multi-group fairness auditing, the disagreement set consists of *pairs* of elements drawn from distinct protected groups, encoding comparative uncertainty rather than pointwise ambiguity. This distinction originates from two structural differences: First, the input space in the auditing setting carries additional organization

through the presence of a protected attribute, which induces a partition into groups. Second, the underlying model classes differ in purpose: in learning, the hypothesis class represents candidate models themselves, whereas in auditing it represents surrogate preference functionals capturing only the discriminative behavior of the fixed black-box model.

Relationship between multi-star number and Daniely-Shwartz dimesnion. The following claim establishes that the multi-star number is upper bounded by the DS dimension. Consequently, when the DS dimension is infinite, passive learning of the probe class is impossible, and hence active learning is impossible as well. Later, we characterize active learnability of probes in the interactive setting induced by CGQs in terms of both these information complexities. In particular, we show that the finiteness of the multi-star number provides a necessary and sufficient condition for learning probes in this interactive setting.

Claim 13. *For an arbitrary multiclass $\mathcal{F}$, the following holds:*

$$\mathfrak{s}_k(\mathcal{F}) \leq \mathrm{DS}_k(\mathcal{F})$$

Proof Let $S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ denotes a DS-shattered set, $B \subseteq \mathcal{F}[S]$ an m-dimensionnal cube, and fix $b_0 \in B$ realizable by $f_0 \in \mathcal{F}$ in S.

for every i in $[m]$, by the pseudo-cube definition, there exists an i-neighbor b_i in B of b.

We have for every $j \in [m] \setminus \{i\}$, $b_i^i \neq b_i^0$ and $b_j^i = b_j^0$

Let f_i in $\mathcal{F}$ a probe that realizes b_i. That is $f_i|_S = b^i$. Since $f_0|_S = b^0$, we obtain

$$f_i(\mathbf{x}_i) = b_i^i \neq b_i^0 = f_0(\mathbf{x}_i),$$

and for every $j \in [m] \setminus \{i\}$,

$$f_i(\mathbf{x}_j) = b_j^i = b_j^0 = f_0(\mathbf{x}_j).$$

Then for every i in $[m]$, $\mathrm{DIS}(\{f_0, f_i\}) \cap S = \{\mathbf{x}_i\}$.

Hence S is also a star set of size m.

By taking the supremum over DS-shattered S, we obtain the desired property.

■

C.1.1 Hardness Results on Sample Complexity and Model Extraction

We begin by establishing a lower bound on the sample complexity required to learn bias probes.

Theorem 14 (Sample Complexity Lower Bound)**.** *Fix $\varepsilon \in (0, 1/4), \delta \in (0, 1/2)$. For any (possibly randomized) active auditor making at most*

$$m \leq \Omega\left(\max\left\{\frac{1-2\delta}{1-\delta}\min\left\{\mathfrak{s}_k, \left\lfloor\frac{1}{2\epsilon}\right\rfloor\right\}, \, [1-2\epsilon-2\delta(1-\epsilon)]_+ \,\mathrm{DS}_k\right\}\right)$$

CGQ queries, there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and a target comparison functional $f_{h^*} \in \mathcal{F}$ such that, with probability at least δ, $\mathop{\mathbb{P}}_{\mathcal{Q}_{1:m}\sim\mathcal{D}^{mk}}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{D}^k}\left[\hat{f}_{\mathcal{Q}_{1:m}}(\mathbf{X}) \neq f_h(\mathbf{X})\right] > \varepsilon\right] > \delta$.

Theorem 14 shows that learning bias probes is fundamentally limited by the complexity of the comparison class $\mathcal{F}$. In particular, even in the active setting, the sample complexity must scale with intrinsic structural parameters such as $\mathfrak{s}_k$. This highlights that auditing through comparison functionals does not circumvent statistical hardness, but rather shifts it to the geometry of the induced comparison class. The proof is given in Appendix E.1.

C.1.2 HARDNESS OF MODEL EXTRACTION ATTACKS

In this section, we characterize the hardness of extracting the model underlying $\mathbf{M}$ from both computational and information-theoretic perspectives.

Computational Hardness. We now show that, beyond the statistical limitations established in Proposition 2, extracting the underlying model from bias probes is computationally hard. Our probe-based auditing procedure induces an interactive protocol in which the auditor interacts with the model owner through at most B cross-group queries (CGQs). We formalize the corresponding model-extraction problem through the following notion of *B-Bounded Relational Extraction*, that consists of:

- Finite hypothesis class $\mathcal{H}_0 = \{h_1, \ldots, h_m\}$,
- Finite collection of admissible relational queries $\mathcal{Q} = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\} \subseteq \mathcal{X}$,
- The query budget of k-ALEBI B.

Model owner's model extraction corresponds to whether there exists a subset of queries $\mathcal{Q}' \subseteq \mathcal{Q}$, such that $|\mathcal{Q}'| \leq B$ and such that the answers to the queries in $\mathcal{Q}'$ uniquely identify the unknown hypothesis in $\mathcal{H}_0$. Equivalently, the question is whether for any two different models h and g in $\mathcal{H}_0$ there exists CGQ queries $\mathcal{Q}'$ that distinguish f_h from f_g. Thus, the goal is to make the map $h \longmapsto \big(f_h(q)\big)_{q \in Q'}$ injective on $\mathcal{H}_0$.

Theorem 15 (Hardness of Model Extraction via Bias Probes)**.** *Given a learned probe $f_h \in \mathcal{F}_{\mathcal{H}}$ from a (black-box) interaction with a model owner $\mathbf{M}$ of $h \in \mathcal{H}$, recovering h is NP hard.*

Theorem 15 establishes a computational barrier to model extraction from bias probes. Recovering the underlying model h from CGQ-based queries is intractable. More precisely, finding a small set of CGQ sufficient for universal model extraction is NP-complete. This formalizes a separation between auditing and reconstruction: while bias probes suffice to estimate properties such as multi-group fairness, they do not reveal enough information to reconstruct the model efficiently.

Proof We reduce from the Minimum Test Cover problem (Definition **??**), known to be NP-complete.

We begin by showing that B-BOUNDED-RELATIONAL-EXTRACTION is NP-hard via a polynomial-time reduction from MINIMUM-TEST-COVER, even when the multi-context class and the set of admissible queries are finite and every query uses two points that occur in no other admissible query.

Membership in NP is immediate: given a candidate $\mathcal{Q}' \subseteq \mathcal{Q}$, one can verify $|\mathcal{Q}'| \leq B$ and, for each of the at most $\binom{m}{2}$ pairs of hypotheses, check whether some query in $\mathcal{Q}'$ distinguishes the pair. By assumption, every relational response is polynomial-time computable. For NP-hardness, we reduce from MINIMUM TEST-COVER.

Let $(U, \mathcal{T}, K)$ be an instance, where $U = \{u_1, \ldots, u_m\}$, $\mathcal{T} = \{T_1, \ldots, T_r\}$. A subcollection $\mathcal{T}' \subseteq \mathcal{T}$ is a test cover if for every pair $u_i \neq u_j$, there exists $T_\ell \in \mathcal{T}'$ containing exactly one of u_i and u_j.

First, we can observe that if some pair $u_i \neq u_j$ is not distinguished by any test in $\mathcal{T}$, then the TEST-COVER instance is necessarily a NO instance. Such instances may therefore be mapped in polynomial time to any fixed NO instance of the target problem. Hence, for the remainder of the reduction, we may assume that every pair of objects is distinguished by at least one test. For every test T_ℓ, introduce two fresh domain points a_ℓ, b_ℓ. All such points are distinct, so $\{a_\ell, b_\ell\} \cap \{a_{\ell'}, b_{\ell'}\} = \varnothing$, for $\ell \neq \ell'$.

For every object $u_j \in U$, define a hypothesis $h_j : \mathcal{X} \to \{0, 1\}$ by $h_j(a_\ell) = 1$ and $h_j(b_\ell) = 1 - \mathbb{1}\{u_j \in T_\ell\}$. Since every pair of objects is distinguished by at least one test, the resulting hypotheses $h_1, \ldots, h_m$ are pairwise distinct.

For every T_ℓ, define the relational query $q_\ell = (a_\ell, b_\ell)$ and let $f_h(q_\ell) = h(a_\ell) - h(b_\ell)$. Then

$$f_{h_j}(q_\ell) = h_j(a_\ell) - h_j(b_\ell) = \mathbb{1}\{u_j \in T_\ell\}.$$

Consequently, for every $i \neq j$,

$$f_{h_i}(q_\ell) \neq f_{h_j}(q_\ell) \iff \mathbb{1}\{u_i \in T_\ell\} \neq \mathbb{1}\{u_j \in T_\ell\}$$
$$\iff T_\ell \text{ distinguishes } u_i \text{ and } u_j.$$

Now set $B = K$. For any subcollection $\mathcal{T}' \subseteq \mathcal{T}$, define $\mathcal{Q}' = \{q_\ell : T_\ell \in \mathcal{T}'\}$. The correspondence $T_\ell \leftrightarrow q_\ell$ is one-to-one, so

$$|\mathcal{Q}'| = |\mathcal{T}'|.$$

Moreover, by the equivalence above,

$$\mathcal{T}' \text{ distinguishes every pair in } U \iff \mathcal{Q}' \text{ distinguishes every pair in } \mathcal{H}_0.$$

Therefore,

$$(U, \mathcal{T}, K) \in \textsc{Test-Cover} \iff (\mathcal{H}_0, \mathcal{Q}, B) \in \textsc{Bounded-Relational-Extraction}.$$

The construction uses $2r$ domain points, m hypotheses, and r queries, and its explicit representation can be constructed in $O(mr)$ time. Hence the reduction is polynomial. Finally, since every q_ℓ uses the fresh pair $\{a_\ell, b_\ell\}$, distinct queries have disjoint supports. Therefore the NP-completeness result holds even under this restriction. ■

D Information-theoretic Characterizaiton of Model Extraction Attacks

Information-Theoretic Hardness. In this section, we quantify the information revealed about the model owner's identity through interactions with the auditor. We assume that $\mathcal{H}$ is finite and that the induced multi-context class relative to $\mathcal{H}$, denoted by $\mathcal{F} = \mathcal{F}_{\mathcal{H}}$, is therefore also finite. We further assume a uniform prior over $\mathcal{H}$.

Before the audit procedure begins, the uncertainty about the model identity is $\widetilde{\mathcal{I}}_0 := \log_2 |\mathcal{H}|$ bits. After t interactions with $\mathbf{M}$, the observed responses induce a version space $\mathcal{V}_t \subseteq \mathcal{H}$ containing the hypotheses consistent with all interactions observed so far. Under the uniform prior, the remaining model identity uncertainty is therefore $\widetilde{\mathcal{I}}_t := \log_2 |\mathcal{V}_t|$. This motivates the following definition of identity leakage.

Definition 8 (Identity Leakage)**.** *For a hypothesis class $\mathcal{H}$, the identity leakage after t interactions between the auditor and the model owner is defined as*

$$\mathcal{I}_t := \widetilde{\mathcal{I}}_0 - \widetilde{\mathcal{I}}_t = \log_2 \frac{|\mathcal{H}|}{|\mathcal{V}_t|}.$$

For example, suppose the auditor knows that the model belongs to a hypothesis class of size $|\mathcal{H}| = 1000$. If, after t interactions with the model owner, only $|\mathcal{V}_t| = 20$ hypotheses remain consistent with the observed responses, then the identity leakage is

$$\mathcal{I}_t = \log_2 \frac{1000}{20} \approx 5.64$$

bits. Thus, the audit interactions have reduced the uncertainty about the model identity by approximately 5.64 bits.

In Section 4, we design an experimental protocol to evaluate model extraction through identity leakage. We compare our proposed auditor with three baselines: ReconAudit, DirectAudit, and the auditing method proposed by (Yan & Zhang, 2022).

E PROOFS FOR MANIPULATION-FREE REGIME

E.1 LOWER BOUND ON SAMPLE COMPLEXITY

Theorem 14 (Sample Complexity Lower Bound). *Fix $\varepsilon \in (0, 1/4), \delta \in (0, 1/2)$. For any (possibly randomized) active auditor making at most*

$$m \leq \Omega\left(\max\left\{\frac{1-2\delta}{1-\delta}\min\left\{\mathfrak{s}_k, \left\lfloor\frac{1}{2\epsilon}\right\rfloor\right\}, \left[1-2\epsilon-2\delta(1-\epsilon)\right]_+ \mathrm{DS}_k\right\}\right)$$

CGQ queries, there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and a target comparison functional $f_{h^*} \in \mathcal{F}$ such that, with probability at least δ, $\mathop{\mathbb{P}}_{\mathcal{Q}_{1:m}\sim\mathcal{D}^{mk}}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{D}^k}\left[\hat{f}_{\mathcal{Q}_{1:m}}(\mathbf{X}) \neq f_h(\mathbf{X})\right] > \varepsilon\right] > \delta$.

Proof

Step 1: Active bound.

Let $\alpha_\epsilon = \min\left\{\mathfrak{s}_k, \left\lfloor\frac{1}{2\epsilon}\right\rfloor\right\}$. Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\alpha_\epsilon}\}$ be a k-star with center f_0 and witnesses $f_1, \ldots, f_{\alpha_\epsilon}$, and define the fixed coupling $\nu = \frac{1}{\alpha_\epsilon}\sum_{i=1}^{\alpha_\epsilon}\delta_{\mathbf{x}_i}$.

We have, $\nu(\{\mathbf{x}_i\}) = \frac{1}{\alpha_\epsilon} \geq 2\epsilon > \epsilon$.

Consider any randomized auditor using at most q queries. We begin by observing :

$$\sup_{f^*\in\mathcal{F}_k}\mathbb{P}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon\right] \geq \max_{i\in[\alpha_\epsilon]}\mathbb{P}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f_i(\mathbf{X})] > \epsilon\right]$$

We proceed by coupling the audit process with targets $f_0, f_1, \ldots, f_{\alpha_\epsilon}$ using the same unlabeled stream and the same internal randomness for the audit.

Let Q_0 denote the distinct star points queried under target f_0, and if

$$\mathbb{P}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_0(\mathbf{X}) \neq f_0(\mathbf{X})] > \epsilon\right] > \delta$$

f_0 is already hard.

Otherwise, for all $i \in [\alpha_\epsilon]$ with probability at least $1-\delta$, we have

$$\widehat{f}_0(\mathbf{x}_i) = f_0(\mathbf{x}_i) \tag{5}$$

Let E_0 denotes that event, that is

$$E_0 = \left\{\widehat{f}_0(\mathbf{x}_i) = f_0(\mathbf{x}_i) : \forall i \in [\alpha_\epsilon]\right\}$$

We have

$$\mathbb{P}\left|E_0\right] \geq 1-\delta \tag{6}$$

If $\mathbf{x}_i \notin Q_0$, then all queries lie in some $\mathbf{x}_j$, where $j \neq i$.

By the star definition, we know that $f_i(\mathbf{x}_j) = f_0(\mathbf{x}_j)$, and a simple induction over the interaction wxith the model owner therefore gives $\widehat{f}_i = \widehat{f}_0$.

Hence, on the event in 5, every non-queried $\mathbf{x}_i$ produces $\widehat{f}_i$ such that $\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_i(\mathbf{X}) \neq f_i(\mathbf{X})] > \epsilon$.

Let F_i denote the event

$$F_i = \left\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_i(\mathbf{X}) \neq f_i(\mathbf{X})]\right\}$$

We have shown that

$$E_0 \cap \{\mathbf{x}_i \in Q_0\} \subseteq F_i$$

Therefore,

$$\mathbb{1}_{E_0}\mathbb{1}_{\mathbf{x}_i \notin Q_0} \leq \mathbb{1}_{F_i} \tag{7}$$

On the other hand, by definition $|Q_0| \leq q$. Since there α_ϵ star points,

$$\sum_{i\in[\alpha_\epsilon]} \mathbb{1}_{\mathbf{x}_i \notin Q_0} \geq \alpha_\epsilon - q \tag{8}$$

$$\begin{aligned}
\max_{i\in[\alpha_\epsilon]} \mathbb{P}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_i(\mathbf{X}) \neq f_i(\mathbf{X})] > \epsilon\right] &\geq \frac{1}{\alpha_\epsilon}\sum_{i\in[\alpha_\epsilon]} \mathbb{P}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_i(\mathbf{X}) \neq f_i(\mathbf{X})] > \epsilon\right] \\
&= \frac{1}{\alpha_\epsilon}\sum_{i\in[\alpha_\epsilon]} \mathbb{E}[\mathbb{1}_{F_i}] \\
&\geq \frac{1}{\alpha_\epsilon}\sum_{i\in[\alpha_\epsilon]} \mathbb{E}[\mathbb{1}_{E_0}\mathbb{1}_{\mathbf{x}_i \notin Q_0}] \\
&\geq \frac{\alpha_\epsilon - q}{\alpha_\epsilon}\,\mathbb{P}[E_0] \\
&\geq \frac{(\alpha_\epsilon - q)(1-\delta)}{\alpha_\epsilon}
\end{aligned}$$

Where step 3 follows from 7, step 4 follows from 8 , and the last step follows from 6.

The right part is bigger than δ when $\frac{\alpha_\epsilon - q}{\alpha_\epsilon}(1-\delta) > \delta$,

which is equivalent to $q > \frac{1-2\delta}{1-\delta}\alpha_\epsilon$.

Step 2: Passive bound. Let $m = \mathrm{DS}_k(\mathcal{F})$, $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ DS-shattered CGQs, and $B \subseteq \mathcal{F}[S]$ an m-dimensional pseudo-cube. Let ν the coupling be defined as: $\nu = \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{x}_i}$.

For the lower bound we strengthen the learner by granting it direct membership query access to any of the m ν-support points. We can note that a lower bound in this stronger model is also a lower bound for the pool-based auditor.

Let the target f^* be uniformly random from B. By Yao's minimax principle (Yao, 1977), it suffices to analyze a deterministic auditor.

Suppose the auditor has made r distinct queries, with $r \leq q$. Condition on an arbitrary resulting interaction history, which we denote T.

Let $B_T = \{\mathbf{b} \in B : b \text{ is consistent with } T\}$. Because the prior on B is uniform and the auditor is deterministic, the conditional target is uniform on B_T.

Fix an non-queried coordinate $i \in [m]$. For every $\mathbf{b} \in B_T$, the pseudo-cube property gives an i-neighbor $\mathbf{b}' \in B$ satisfying $b'_i \neq b_i$, and $b'_j = b_j$ for $j \neq i$ and since coordinate i has not been queried, $\mathbf{b}'$ agrees with $\mathbf{b}$ on every queried coordinate hence $\mathbf{b}' \in B_T$.

For $\mathbf{b} = (b_1, b_2, \cdots, b_{i-1}, b_i, b_{i+1}, \cdots, b_m)$, let $\pi_i(\mathbf{b}) = (b_1, b_2, \cdots, b_{i-1}, b_{i+1}, \cdots, b_m)$ denote the projection removing the ith coordinate from $\mathbf{b}$. For $g \in \pi_i(B_T)$, we define a fiber $F_g = \{\mathbf{b} \in B_T : \pi_i(\mathbf{b}) = g\}$

Partition B_T into these fibers. Every such fiber contains at least two traces and within one fiber, all labels at coordinate i are distinct; if two traces had the same value at coordinate i and agreed on all other coordinates, they would be identical.

Therefore, for any label $\mathbf{y}_i$ predicted by the auditor at $\mathbf{x}_i$, in every fiber at most one trace has $b_i = \mathbf{y}_i$.

By definition of the pseudo-cube, every fiber has cardinality at least two and since conditioned on T , the target f^* is uniform on B_T, it is also uniform within each fiber conditioned on membership of that fiber, by total law of probability we obtain

$$\mathbb{P}\left[\widehat{f}(\mathbf{x}_i) \neq f^*(\mathbf{x}_i) \mid T\right] \geq \frac{1}{2}.$$

There are at least $m - r \geq m - q$ non-qeried coordinates. Hence

$$\begin{aligned}\mathbb{E}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X}) \mid T\right] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left(\widehat{f}(\mathbf{x}_i) \neq f^*(\mathbf{x}_i) \mid T\right) \\ &\geq \frac{m-q}{2m}.\end{aligned}$$

Taking expectation over history,

$$\mathbb{E}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})\right] \geq \frac{m-q}{2m}. \tag{9}$$

Now suppose the auditor were (ϵ, δ)-PAC for every target in the pseudo-cube. Averaging over the uniform prior on B gives

$$\mathbb{P}\left\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon\right\} \leq \delta$$

Since

$$0 \leq \mathbb{P}\left\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon\right\} \leq \delta) \leq 1,$$

we then have

$$\begin{aligned}\mathbb{E}[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})]] &\leq \epsilon\,\mathbb{P}\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})] \leq \epsilon\} + \mathbb{P}\{\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon\} \\ &\leq \epsilon(1-\delta) + \delta \\ &= \epsilon + \delta(1-\epsilon).\end{aligned}$$

Combining with 9,

$$\frac{m-q}{2m} \leq \epsilon + \delta(1-\epsilon).$$

Therefore

$$q \geq m\left[1 - 2\epsilon - 2\delta(1-\epsilon)\right]$$

■

E.2 Proof of ALEBI Upper Bounds

Theorem 3 (Upper bounds of ALEBI). *Algorithm 1 is a PAC active auditor: with probability at least* $1-\delta$*, it outputs* $\hat{f}$ *satisfying* $\mathbb{P}\left\{\hat{f}(Z) \neq f_{h^*}(Z)\right\} \leq \epsilon$ *using at most*

$$\mathcal{O}\left(\min\left\{n,\ (e-1)\min\{\mathfrak{s}_k, n\}\log\frac{en}{\min\{\mathfrak{s}_k,n\}} + \log\frac{2}{\delta} + 1,\ \frac{\min\{\mathfrak{s}_k,n\}}{1\vee\log\min\{\mathfrak{s}_k,n\}}\,\mathrm{DS}_k\log\left(\frac{en(2^k-2)}{\mathrm{DS}_k}\right)\right\}\right)$$

queries. where $n = \left\lceil\frac{\mathrm{DS}_k+\log(4/\delta)}{\epsilon}\right\rceil$.

Proof We organize the proof as follows: we begin by showing the active terms in the upper bound by adapting the proof of Hanneke (2014) to the multi-class learning problem, since the probe learning problem reduces to multiclass learning where the size of labels depend on the cardinal of groups k.

Step 1: The passive term. Here, the passive learner is allocated failure probability $\delta/2$. Fix $n = \left\lceil\frac{\mathrm{DS}_k+\log(4/\delta)}{\epsilon}\right\rceil$.

By Theorem 20, the passive term derives directly by drawing n CGQ queries $\mathcal{Q}_{1:m} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ from the fixed coupling ν and obtaining their probe labels via $\mathbf{M}$, which guarantees:

By the choice of n, the multiclass passive algorithm proposed in **?** outputs $\hat{f}_{\mathcal{Q}_{1:n}}$ such that:

$$\mathop{\mathbb{P}}_{\mathcal{Q}_{1:n}}\left[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}\left[\widehat{f}_{\mathcal{Q}_{1:n}}(\mathbf{X}) \neq f^*(\mathbf{X})\right] > \epsilon\right] \leq \frac{\delta}{2} < \delta$$

In the following, fix $S \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ as the sample set used for passive learning, with $n = \left\lceil\frac{\mathrm{DS}_k+\log(4/\delta)}{\epsilon}\right\rceil$.

Step 2: First active term. We assume that pool of CGQ queries is index by some set I. If $I \subseteq [N]$ denotes the set of indices that $\mathbf{A}$ have seen in the history of interaction with $\mathbf{M}$, the corresponding version space is given by $\mathcal{V}_I = \{f \in \mathcal{F}_k : f(\mathbf{x}_j) = f^\star(\mathbf{x}_j), \forall j \in I\}$, where $f^\star = f_{h^\star}$, $h^\star$ being the black-box model under audit.

For a nonempty set I, we say that $i \in I$ is an effective (informative) index if it exhibits disagreement within the multi-context class with respect to $f^\star$, meaning there exists $f_i \in \mathcal{F}_k$ such that for all $j \in I \setminus \{i\}$, $f_i(\mathbf{x}_j) = f^\star(\mathbf{x}_j)$ while $f_i(\mathbf{x}_i) \neq f^\star(\mathbf{x}_i)$. Let $E(I)$ denote such set of effective coordinates.

It is straightforward to see that $|E(I)| \leq \min\{\mathfrak{s}_k, |I|\}$

Let $f_0, \{f_i\}_{i\in I}$ be such that they exhibit this disagreement within $\mathcal{F}_k$, where $f_0 = f^\star$

For $i \neq j \in E(I)$, we have $f_i(\mathbf{x}_j) = f^\star(\mathbf{x}_j)$ while $f_i(\mathbf{x}_i) \neq f^\star(\mathbf{x}_i)$

If $\mathbf{x}_i = \mathbf{x}_j$ for $i \neq j$, then a witness for coordinate i would have to both agree and disagree with $f^\star$ at the same CGQ. Therefore the CGQs indexed by the effective set $E(I)$ must be distinct.

which shows that $\{\mathbf{x}_i : i \in E(I)\}$ is a multi-star set for $\mathcal{F}_k$ centered at $f_0 = f^\star$. By definition of $\mathfrak{s}_k$, $|E(I)| \leq \mathfrak{s}_k$.

For an integer n [5], let $s_n \triangleq \min\{\mathfrak{s}_k, n\}$. For $I \subseteq [n]$, let $N(I)$ be the cardinal of CGQ queries when the coordinates of I are processed in a uniformly random order, and let $\Psi(I) \triangleq \mathbb{E}\left[e^{\lambda N(I)} \,\middle|\, S\right]$, where $\lambda \geq 0$.

We show by induction on $m = |I|$ that the following holds

$$\Psi(I) \leq \prod_{r=1}^{m}\left[1 + (e^\lambda - 1)\min\left\{1, \frac{s_n}{r}\right\}\right]$$

For $m = 0$ the resulting identity is trivial.

[5] n here is arbitrary, we'll later use n as the sample complexity for passive learning probes.

Let $|I| = m \geq 1$. In a uniformly random ordering of I, the last coordinate is uniformly distributed over I.

Conditionned on coordinate i being last. After all coordinates in $I \setminus \{i\}$ have been processed, the current version space is

$$\{f : f(\mathbf{x}_j) = f^\star(\mathbf{x}_j)\ \forall j \in I \setminus \{i\}\}$$

Therefore the final CGQ cooreponding to index i is queried if and only if $i \in E(I)$. Hence

$$N(I) = N(I \setminus \{i\}) + \mathbb{1}\{i \in E(I)\}$$

Averaging over the uniformly random last coordinate yields

$$\Psi(I) = \frac{1}{m} \sum_{i \in I} e^{\lambda \mathbf{1}\{i \in E(I)\}} \Psi(I \setminus \{i\}).$$

Applying the induction hypothesis to each $I \setminus \{i\}$ yields

$$\Psi(I) \leq \Big(\prod_{r=1}^{m-1} \Big[1 + (e^\lambda - 1) \min\Big\{1, \frac{s_n}{r}\Big\}\Big] \Big) \Big(\frac{1}{m} \sum_{i \in I} e^{\lambda \mathbf{1}\{i \in E(I)\}} \Big)$$

On the other hand,

$$\frac{1}{m} \sum_{i \in I} e^{\lambda \mathbf{1}\{i \in E(I)\}} = 1 + (e^\lambda - 1) \frac{|E(I)|}{m}$$

and since $|E(I)| \leq \min\{\mathfrak{s}_k, |I|\}$, we have $\frac{|E(I)|}{m} \leq \min\big\{1, \frac{s_n}{m}\big\}$ we conclude the induction.

For $I = [n]$ with the fact that $1 + t \leq e^t$, we have

$$\mathbb{E}\left[e^{\lambda N_n} \,\middle|\, S\right] \leq \exp\left((e^\lambda - 1) \sum_{r=1}^{n} \min\Big\{1, \frac{s_n}{r}\Big\} \right),$$

where $N_n \triangleq N([n])$.

For $\lambda = 1$, and by Markov's inequality,

$$\begin{aligned} \mathbb{P}\left\{N_n > b \,|\, S\right\} &= \mathbb{P}\left\{e^{N_n} > e^b \,\middle|\, S\right\} \\ &\leq \exp\Big((e-1) \min\Big\{1, \frac{s_n}{r}\Big\} - b \Big). \end{aligned}$$

Taking $b = (e-1) \min\big\{1, \frac{s_n}{r}\big\} + \log \frac{2}{\delta}$, for every realized pool S,

$$\mathbb{P}\left\{N_n > b \,|\, S\right\} \leq \frac{\delta}{2}$$

The auditor $\mathbf{A}$ runs k-ALeBi for a budget of b under the condition that if it is about to make query number $b + 1$, it halts. Thus the truncated procedure makes at most b oracle queries on every execution.

If truncation does not occur, k-ALeBi has reconstructed every response in the original i.i.d. sample S used for the probe passive leaning, and hence has recovered exactly the target probe. The passive term in **Step 1** is then applied to this completed labeled sample and requires no additional oracle queries.

The failure events are therefore $Z_{\text{alebi}} = \{k\text{-ALeBi is truncated}\}$ for interactive setting

and $Z_{\text{pass}} = \left\{ \mathbb{P}_{\mathbf{X} \sim \nu} [\widehat{f}_S(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon \right\}$ for passive setting.

In step 1, we have shown that $\mathbb{P}(Z_{\text{pass}}) \leq \frac{\delta}{2}$ And in step 2, we have shown that: $\mathbb{P}(Z_{\text{alebi}}) \leq \frac{\delta}{2}$ Therefore, by the union bound,

$$\mathop{\mathbb{P}}_{\mathcal{Q}_{1:n}}[\mathop{\mathbb{P}}_{\mathbf{X}\sim\nu}[\widehat{f}_{\mathcal{Q}_{1:n}}(\mathbf{X}) \neq f^*(\mathbf{X})] > \epsilon] < \delta$$

Finally $1 \leq s_n \leq n$, then

$$\begin{aligned}\sum_{r=1}^{n} \min\left\{1, \frac{s_n}{r}\right\} &= \sum_{m=1}^{s_n} 1 + \sum_{m=s_n+1}^{N} \frac{s_n}{m} \\ &\leq s_n + s_n \log \frac{n}{s_n} \\ &= s_n \log \frac{en}{s_n}.\end{aligned}$$

In particular, for $\mathfrak{s}_k \leq n$, this yields a sample complexityb of

$$(e-1)\mathfrak{s}_k \log \frac{en}{\mathfrak{s}_k} + \log \frac{2}{\delta} + 1 \tag{10}$$

with $n = \left\lceil \frac{\mathrm{DS}_k + \log(4/\delta)}{\epsilon} \right\rceil$.

Step 3: Second active term. We next obtain a deterministic exact-completion procedure on the same finite pool using the same fixed coupling over groups marginals. Let $\mathcal{V}_S = \{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) : f \in \mathcal{F}_k\}$ denote the version space consistent with S.

Recall the definitions of specifying sets and extended teaching dimension in DEfinition 10. For an arbitrary labeling $g \in \mathcal{Y}_k^n$, S is a specifying set for g with respect to $\mathcal{V}_S$ if $|\{v \in \mathcal{V}_S : v|_S = g|_S\}| \leq 1$. The extended teaching dimension is

$$\mathrm{XTD}(\mathcal{V}_S) = \max_{g \in \mathcal{Y}_k^n} \min\{|S| : S \text{ specifies } g \text{ with respect to } \mathcal{V}_S\}.$$

We first establish the relation between XTD and the multiclass star number that is needed in the algorithm:

$$\mathrm{XTD}(\mathcal{V}_S) \leq \min\{\mathfrak{s}_k(\mathcal{V}_S) + 1, n\} \leq \min\{\mathfrak{s}_k + 1, n\}$$

The bound by n is trivial since all n coordinates specify any labeling with respect to $\mathcal{V}_S$.

To prove the first inequality, fix an arbitrary $g \in \mathcal{Y}_k^n$ and let $S = \{i_1, \ldots, i_m\}$ be a specifying set of minimum cardinality.

If $m = 0$ the bound terms become trivial. For every $r \in [m]$, minimality implies that $S \setminus \{i_r\}$ is not specifying. Consequently there exist at least two distinct traces in $\mathcal{V}_S$ agreeing with g on $S \setminus \{i_r\}$.

Because S itself is specifying, at most one of those traces can also agree with g at i_r. We can therefore choose $v_r \in \mathcal{V}_S$ such that $v_r(i_r) \neq g(i_r)$ and for all $\ell \neq r$, $v_r(i_\ell) = g(i_\ell)$.

If $m \geq 2$, let v_1 as a center and the $m-1$ indices $i_2, \ldots, i_m$.

For every $r \geq 2$, $v_r(i_r) \neq g(i_r) = v_1(i_r)$, while for every $\ell \geq 2$ with $\ell \neq r$, $v_r(i_\ell) = g(i_\ell) = v_1(i_\ell)$.

Hence $i_2, \ldots, i_m$ is a multi-star set for $\mathcal{V}_S$, centered at v_1, and consequently, $m - 1 \leq \mathfrak{s}(\mathcal{V}_S)$, Therefore $m \leq \mathfrak{s}(\mathcal{V}_S) + 1$.

The same inequality is trivial for $m = 1$. Since g was arbitrary $\mathrm{XTD}(\mathcal{V}_S) \leq \mathfrak{s}(\mathcal{V}_S) + 1$. Finally, a star in $\mathcal{V}_S$ lifts to a star on the corresponding CGQs in $\mathcal{F}_k$, and therefore $\mathfrak{s}(\mathcal{V}_S) \leq \mathfrak{s}_k$

This proves

$$\mathrm{XTD}(\mathcal{V}_S) \le \min\{\mathfrak{s}_k + 1, n\}$$

Algorithm 4 PROBES MEMBERSHIP-HALVING

Require: Budget m, finite trace $\mathcal{V}_S$.

1: Intialization:
$$\mathcal{V} \leftarrow \mathcal{V}_S.$$
2: **while** $|V| > 1$ **do**
3: **for** $i = 1, \dots, m$ **do**
4: Choose a maximizing response
$$\phi_{\mathcal{V}}(i) \in \arg\max_{y \in \mathcal{Y}_k} |\{v \in \mathcal{V} : v(i) = y\}|.$$
5:
$$a_i := |\{v \in \mathcal{V} : v(i) = \phi_{\mathcal{V}}(i)\}|.$$
6: **end for**
7: Choose a smallest specifying set S for $\phi_{\mathcal{V}}$ with respect to V.
8: Choose
$$i^\star \in \arg\max_{i \in S}(|V| - a_i).$$
9: Query
$$y^\star \leftarrow f^\star(\mathbf{X}_{i^\star}).$$
10: Update
$$\mathcal{V} \leftarrow \{v \in \mathcal{V} : v(i^\star) = y^\star\}.$$
11: **end while**
12: Let $v^\star$ denote the unique surviving trace in $\mathcal{V}$.
13: **return**
$$((\mathbf{X}_i, v^\star(i)))_{i=1}^n.$$

Next we show that Algorithm 4 uses at most $m = 1 + \mathrm{XTD}(\mathcal{V}_S) \log |\mathcal{V}_S|$ CGQs.

Let $V \subseteq \mathcal{V}_S$ denote the current trace version space. With the initialization step $\mathcal{V} = \mathcal{V}_S$, for every index $i \in [n]$, Algorithm 4 chooses a maximizing response $\phi_{\mathcal{V}}(i) \in \arg\max_{y \in \mathcal{Y}_k} |\{v \in \mathcal{V} : v(i) = y\}|$ and let $a_i = |\{v \in \mathcal{V} : v(i) = \phi_{\mathcal{V}}(i)\}|$.

Since restricting a class cannot increase XTD, $\mathrm{XTD}(\mathcal{V}) \le \mathrm{XTD}(\mathcal{V}_S)$.

Hence there exists a set $S \subseteq [n]$ such that $|S| \le \mathrm{XTD}(\mathcal{V}_S)$ that specifies the labeling $\phi_{\mathcal{V}}$.

At most one trace in V agrees with $\phi_{\mathcal{V}}$ on all coordinates in S thus every other trace disagrees with $\phi_{\mathcal{V}}$ somewhere in S, yielding $\sum_{i \in S}(|\mathcal{V}| - a_i) \ge |\mathcal{V}| - 1$

Thus there exists $i^\star \in S$ such that

$$|\mathcal{V}| - a_{i^\star} \ge \frac{|\mathcal{V}| - 1}{\mathrm{XTD}(\mathcal{V}_S)}. \tag{11}$$

Algorithm 4 queries $y^\star = f^\star(\mathbf{x}_{i^\star})$ and retain $\mathcal{V}' = \{v \in \mathcal{V} : v(i^\star) = y^\star\}$.

We distinguish between two cases;

If $y^\star = \phi_{\mathcal{V}}(i^\star)$, then $|\mathcal{V}'| = a_{i^\star}$ and inequality 11 gives

$$|\mathcal{V}'| - 1 \le \left(1 - \frac{1}{\mathrm{XTD}(\mathcal{V}_S)}\right)(|\mathcal{V}| - 1)$$

On the other hand, if $y^\star \ne \phi_{\mathcal{V}}(i^\star)$ then the surviving response cell isn't a maximizing cell. Since a maximizing cell has cardinality at least that of every other cell $|\mathcal{V}'| \le \frac{|\mathcal{V}|}{2}$

When $\mathrm{XTD}(\mathcal{V}_S) \geq 2$,

$$\frac{|\mathcal{V}|}{2} - 1 \leq \left(1 - \frac{1}{\mathrm{XTD}(\mathcal{V}_S)}\right)(|\mathcal{V}| - 1)$$

If $\mathrm{XTD}(\mathcal{V}_S) = 1$, one specifying coordinate leaves at most one surviving trace, so exact identification occurs after one query.

Thus, whenever $|\mathcal{V}| > 1$,

$$|\mathcal{V}'| - 1 \leq \left(1 - \frac{1}{\mathrm{XTD}(\mathcal{V}_S)}\right)(|V| - 1)$$

After q queries,

$$|\mathcal{V}_q| - 1 \leq \left(1 - \frac{1}{\mathrm{XTD}(\mathcal{V}_S)}\right)^q (|\mathcal{V}_S| - 1) \leq e^{-q/\,\mathrm{XTD}(\mathcal{V}_S)}(|\mathcal{V}_S| - 1)$$

If $q > \mathrm{XTD}(\mathcal{V}_S)\log(|\mathcal{V}_S| - 1)$, then the right-hand side is strictly smaller than 1. Since $|\mathcal{V}_q| - 1$ is a nonnegative integer, it must equal zero. Therefore $\mathcal{V}_q$ contains exactly one trace.

Since the target trace is maintained via consistency with version space, that unique trace is exactly $(f^\star(\mathbf{x}_1), \ldots, f^\star(\mathbf{x}_n))$.

Hence the sample complexity of Algorithm 4 is bounded by $1 + \mathrm{XTD}(\mathcal{V}_S)\log|\mathcal{V}_S|$, and since$\mathrm{XTD}(\mathcal{V}_S) \leq \min\{\mathfrak{s}_k+1, n\}$, the sample complexity is bounded by $1+\min\{\mathfrak{s}_k+1, n\}\log|\mathcal{V}_S|$

If $\Pi_{\mathcal{F}_k}(n)$ denotes the growth function over n CGQs, using the fact that $|\mathcal{V}_S| \leq \Pi_{\mathcal{F}_k}(n)$, uniformly over every possible pool, we get a sample complecity of

$$\mathcal{O}(1 + \min\{\mathfrak{s}_k + 1, n\}\log\Pi_{\mathcal{F}_k}(n)) \tag{12}$$

ALgorithm 4 has no truncation failure as it reconstructs the CGQ pool exactly. Thus the only failure event is that of the passive multiclass learner whose probability is already at most $\delta/2$ from **step 1**.

A sharper result presented below is derived to improve the dependence on the extended teaching dimension by a logarithmic factor.

Lemma 16 (Refined bounds for Algorithm 4)**.** *For every finite trace class $\mathcal{V} \subseteq \mathcal{Y}^n$ there exists a deterministic query procedure that exactly identifies every target $v^\star \in \mathcal{V}$ using $\mathcal{O}\Big(\frac{\mathrm{XTD}(\mathcal{V})}{1\vee\log\mathrm{XTD}(\mathcal{V})}\log|\mathcal{V}|\Big)$ queries.*

Proof

For arbitrary label spaces $\mathrm{XTD}(\mathcal{V}_S) \leq \mathfrak{s}(\mathcal{V}_S)$ and restricting on consitency with S doesn't increase the star number $\mathfrak{s}(\mathcal{V}_S) \leq s_k$. Since $\mathrm{XTD}(\mathcal{V}_S) \leq n$, then $\mathrm{XTD}(\mathcal{V}_S) \leq \bar{s}_k$.

We now apply the refined multiclass Membership Halving procedure. At the beginning of interaction, let $\mathcal{V}$ be the current version space and let $\phi_{\mathcal{V}}(i) \in \arg\max_y |\{v \in \mathcal{V} : v(i) = y\}|$ be its maximizing labeling.

Fix a specifying set S for $\phi_{\mathcal{V}}$ of size at most $\mathrm{XTD}(\mathcal{V}_S)$, query the remaining coordinate maximizing $|\{v \in V : v(i) \neq \phi_{\mathcal{V}}(i)\}|$, keeping $\phi_{\mathcal{V}}$ fixed throughout the phase.

Suppose a phase starts with n traces and the first contradiction to the fixed plurality labeling occurs on its q-th query.

Let A_j be the set of traces eliminated by a maximizing-consistent answer at the j-th queried coordinate.

The sets $A_1, \ldots, A_q$ are pairwise disjoint, and by construction of algorithm 4, we end up constraining a chain $|A_1| \geq |A_2| \geq \cdots \geq |A_q|$. Therefore $q|A_q| \leq \sum_{j=1}^{q}|A_j| \leq n$.

After the contradiction, the surviving exact response cell is a subset of A_q, therefore $|\mathcal{V}'| \leq \frac{n}{q}$.

If no contradiction occurs on the entire specifying set, the target trace is uniquely identified.

Suppose algorithm 4 performs P phases. For phase p, let n_{p-1} be the size of the version space at the beginning of the phase, let n_p be its size at the end of the phase, and let q_p be the number of queries made during that phase.Then there exists a universal constant $c > 0$ such that

$$\log \frac{n_{p-1}}{n_p} \geq c \operatorname{Log}(q_p), \tag{13}$$

where $\operatorname{Log}(t) \triangleq 1 \vee \log t$.

Moreover, since the phase queries only coordinates from a specifying set,

$$q_p \leq \operatorname{XTD}(\mathcal{V}_{p-1}) \leq \operatorname{XTD}(\mathcal{V}_S)$$

The function $t \longmapsto \frac{t}{\operatorname{Log}(t)}$ is nondecreasing for $t \geq 1$. Hence, from $q_p \leq \operatorname{XTD}(\mathcal{V}_S)$,

$$\frac{q_p}{\operatorname{Log}(q_p)} \leq \frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log}(\operatorname{XTD}(\mathcal{V}_S))}.$$

Equivalently,

$$q_p \leq \frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log}(\operatorname{XTD}(\mathcal{V}_S))} \operatorname{Log}(q_p).$$

With the result of 13,

$$q_p \leq \frac{1}{c} \frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log}(\operatorname{XTD}(\mathcal{V}_S))} \log \frac{n_{p-1}}{n_p}.$$

Summing over all P phases gives sample complexity of

$$\begin{aligned} \sum_{p=1}^{P} q_p &\leq \frac{1}{c} \frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log}(\operatorname{XTD}(\mathcal{V}_S))} \sum_{p=1}^{P} \log \frac{n_{p-1}}{n_p} \\ &= \frac{1}{c} \frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log}(\operatorname{XTD}(\mathcal{V}_S))} \log \frac{n_0}{n_P}. \end{aligned}$$

The algorithm starts from $n_0 = |\mathcal{V}_S|$ and terminates when a unique trace remains, that is $n_P = 1$. Therefore a sample complexity of $\frac{\operatorname{XTD}(\mathcal{V}_S)}{\operatorname{Log} \operatorname{XTD}(\mathcal{V}_S)} \log |\mathcal{V}_S|$ suffices

■

Applying lemma 16 with $|\mathcal{V}_S| \leq \Pi_{\mathcal{F}_k}(n)$, and therefore a set of size $O\Big(\frac{\min\{\mathfrak{s}_k, n\}}{1 \vee \log \min\{\mathfrak{s}_k, n\}} \log \Pi_{\mathcal{F}_k}(n)\Big)$ suffices.

On the other hand, a binary point $v = (v_1, \ldots, v_k)$ from the cube $\{0,1\}^k$ induces probe response $T(v) = (v_i - v_j)_{i<j}$

If for two points v, w, one has $T(v) = T(w)$, then for all $i < j$, $v_i - w_i = v_j - w_j$, making $v - w$ constant. The only distinct pair with this property is $(v, w) = (0^k, 1^k)$. Hence the CGQ response alphabet has size $2^k - 1$.

We have $\operatorname{DS}_k(\mathcal{V}_S) \leq \min\{\operatorname{DS}_k(\mathcal{F}), n\}$, applying the multiclass Sauer lemma to the finite class $\mathcal{V}_S \subseteq [2^k - 1]^N$ yields

$$|\mathcal{V}_S| \leq \sum_{j=0}^{\min\{\operatorname{DS}_k, n\}} \binom{n}{j} (2^k - 1 - 1)^j = \sum_{j=0}^{\min\{\operatorname{DS}_k, n\}} \binom{n}{j} (2^k - 2)^j.$$

Consequently, for $\min\{\operatorname{DS}_k, n\} \geq 1$,

$$\log |\mathcal{V}_S| \leq \min\{\operatorname{DS}_k, n\} \log \left(\frac{en(2^k - 2)}{\min\{\operatorname{DS}_k, n\}} \right)$$

Implement this result over 12 gives the sample complexity of

$$\min\{\mathfrak{s}_k, n\} \min\{\mathrm{DS}_k, n\} \log\left(\frac{en(2^k-2)}{\min\{\mathrm{DS}_k, n\}}\right)$$

With $n = \left\lceil \frac{\mathrm{DS}_k + \log(4/\delta)}{\epsilon} \right\rceil$.

Combining the bounds derived in step1,2 and 3 yields the result.

■

For fixed sufficiently small δ, the lower bound derived in Theorem 14 becomes $\max\{\mathrm{DS}_k, \min\{\mathfrak{s}_k, 1/\epsilon.\}\}$, while the upper bound becomes $\min\left\{\frac{\mathrm{DS}_k}{\epsilon},\ \min\{\mathfrak{s}_k, \mathrm{DS}_k/\epsilon\} \log \frac{e\mathrm{DS}_k/\epsilon}{\min\{\mathfrak{s}_k, \mathrm{DS}_k/\epsilon\}},\ \frac{\min\{\mathfrak{s}_k, \mathrm{DS}_k/\epsilon\}\mathrm{DS}_k}{1 \vee \log \min\{\mathfrak{s}_k, \mathrm{DS}_k/\epsilon\}} \log bigl(\frac{e(2^k-2)}{\epsilon}\bigr)\right\}$. If m denotes the sample complexity, and if $\epsilon \lesssim 1/\mathfrak{s}_k$, then $\mathfrak{s}_k \lesssim m \lesssim \mathfrak{s}_k \log \frac{e\mathrm{DS}_k}{\epsilon \mathfrak{s}_k}$, and the bounds match up to logarithmic factors. If $1/\mathfrak{s}_k \lesssim \epsilon \lesssim 1/\mathrm{DS}_k$, then $\frac{1}{\epsilon} \lesssim m \lesssim \frac{\mathrm{DS}_k}{\epsilon}$, with a smaller logarithmic gap whenever $\epsilon \mathfrak{s}_k = \mathcal{O}(1)$. Finally, if $\epsilon \gtrsim 1/\mathrm{DS}_k$, then $\mathrm{DS}_k \lesssim m \lesssim \mathrm{DS}_k/\epsilon$,. In particular m reduces to $\Theta(\mathrm{DS}_k)$ for constant accuracy. Thus the only potentially substantial gap occurs in the intermediate regime and is controlled by the separation between s_k and D_k; in particular, if $D_k = O(1)$ or $\mathfrak{s}_k = \Theta(\mathrm{DS}_k)$, the upper and lower bounds match up to logarithmic factors throughout the accuracy range. Theorem 3 further reveals two distinct regimes. When $\mathfrak{s}_k = \infty$, the active learning component of the bound vanishes, and the complexity is governed by the passive learning term $\mathrm{DS}_k(\mathcal{F})$. In this regime, the auditing problem effectively reduces to passive learning of the probe class, as established by Claim 13. The proof of Theorem 3 is given in Appendix E.2. The following corollary characterizes the learnability of the probe class in the interactive setting in terms of the multi-star number.

F PROOFS FOR FAIRNESS-AWARE MANIPULATION REGIME

Algorithm 5 ROBUST k-ALEBI

Require: k-ALEBI, attack bound p, fixed cell map $q \mapsto J(q)$

if $p = 0$ **then**
  Run subroutine k-ALEBI
  Whenever k-ALEBI requests CGQ q_t, query q_t directly and return the clean response to k-ALEBI
  return the output of k-ALEBI
end if
Configure k-ALEBI with confidence $\delta/2$
$N \leftarrow N_0(\epsilon, \delta/2)$

$$R \leftarrow \left\lceil \frac{2}{(1-2\beta)^2} \log \frac{2N}{\delta} \right\rceil$$

while k-ALEBI has not terminated **do**
  Let q_t be the next CGQ requested by k-ALEBI

$$\widehat{Y}_t \leftarrow \text{LOCALLABEL}(q_t, J(q_t), R)$$

  Feed $\widehat{Y}_t$ to k-ALEBI as the simulated clean response $f^\star(q_t)$
end while
return the output of k-ALEBI

Subroutine LOCALLABEL

Input: CGQ q, fixed cell grid $J(q)$, local budget R
for $r = 1, \ldots, R$ **do**
  Sample independently

$$q'_r \sim \mathcal{D}_Q(\cdot \mid J(q))$$

  Query the attacked oracle at q'_r and receive $\widetilde{Y}_r \in \mathcal{Y}_k$
end for
return

$$\widehat{Y}(q) \in \arg\max_{y \in \mathcal{Y}_k} \sum_{r=1}^{R} \mathbf{1}\{\widetilde{Y}_r = y\}$$

F.0.1 GENERAL INFORMATION-THEORETIC HARDNESS

We begin by introducing a restricted information complexity,

Definition 9 (Fairness-effective star number)**.** *The fairness-effective star number $\mathfrak{s}_\mu(\mathcal{F})$ is the supremum of all integers m for which there exist distinct CGQs $q_1, \ldots, q_m$ and functions $f_0, f_1, \ldots, f_m \in \mathcal{F}$ such that for every $j \in [m]$ $f_0(q_j) = \mathbf{0}$ and, for every $i \in [m]$ $f_i(q_i) \neq \mathbf{0}$. and $j \neq i$ $f_i(q_j) = \mathbf{0}$.*

In the following we consider the assumption:

Assumption 17 (Cell-replicability on the star set)**.** *The fairness-effective star set has pairwise disjoint measurable cells $J_1, \ldots, j_m$ such that, for every witness f_i and every cell J_j, f_i is constant on J_j and equals its value at q_j.*

Theorem 18 (Fairness lower bound under probabilistic attack)**.** *Under assumption 17 and $0 < p < 1/2$. A set of CGQs of size $\Omega\big(\min\{\mathfrak{s}_\mu(\mathcal{F}), \frac{1}{\varepsilon}\}\big)$ is necessary to learn the undelrying probe.*

In Section 4, we evaluate ROBUST k-ALEBI and compare its performance with k-ALEBI, RECONAUDIT and the auditing method proposed by (Yan & Zhang, 2022) in the fairness-aware adversarial regime.

We begin by showing the following lemma

Lemma 19 (Accuracy of one neighborhood simulation)**.** *Under Assumptions 5 and 6,*

$$\mathop{\mathbb{P}}_{q\sim\nu}\Big(\widehat{f}(q)\neq f^\star(q)\Big)\leq \exp\!\left(-\frac{R(1-2\beta)^2}{2}\right).$$

Proof Fix q a CGQ query and define $Z_r \triangleq \mathbb{1}\{\widetilde{Y}_r \neq f^\star(q)\}$.

Step 1: bound the conditional error probability. Let $m_r \triangleq P(f^\star(q'_r)\neq f^\star(q)\,|\,\mathbb{H}_{r-1})$ Since q'_r is a fresh draw from $P_Q(\cdot\mid J(q))$, Assumption 5 gives $m_r\leq\rho$.

If the clean local response differs from $f^\star(q)$, the prob error is bounded by 1. Otherwise an error can occur only through an attack, whose conditional probability is at most p. Hence

$$\begin{aligned}\mathbb{E}[Z_r\mid\mathbb{H}_{r-1}] &\leq m_r+(1-m_r)p\\ &\leq \rho+(1-\rho)p=\beta\end{aligned}$$

Step 2: Failure mode vs the number of errors. If $\sum_{r=1}^R Z_r<\frac{R}{2}$ then strictly more than half of the responses equal $f^\star(q)$. Therefore $f^\star(q)$ is the unique mode. Thus

$$\{\widehat{Y}(q)\neq f^\star(q)\}\subseteq\left\{\sum_{r=1}^R Z_r\geq\frac{R}{2}\right\}$$

Step 3: Concentration of bad events. Fix $\mu_r:=\mathbb{E}[Z_r\mid\mathbb{H}_{r-1}]$ and $D_r:=Z_r-\mu_r$. We have

$$\mathbb{E}[D_r\mid\mathbb{H}_{r-1}]=0,$$

and, conditionally on $\mathbb{H}_{r-1}$, D_r lies in an interval of length one. The conditional Hoeffding lemma therefore gives

$$\mathbb{E}\big[e^{\lambda D_r}\,\big|\,\mathbb{H}_{r-1}\big]\leq e^{\lambda^2/8}.$$

Iterating this bound and applying Chernoff yields

$$\mathbb{P}\left(\sum_{r=1}^R D_r\geq t\right)\leq\exp\!\left(-\frac{2t^2}{R}\right)\tag{14}$$

Since $\mu_r\leq\beta$,

$$\left\{\sum_{r=1}^R Z_r\geq\frac{R}{2}\right\}\subseteq\left\{\sum_{r=1}^R D_r\geq R\left(\frac{1}{2}-\beta\right)\right\}.$$

Replacing $t=R\left(\frac{1}{2}-\beta\right)$ in 14 yields

$$\mathbb{P}\left(\sum_{r=1}^R Z_r\geq\frac{R}{2}\right)\leq\exp\!\left(-\frac{R(1-2\beta)^2}{2}\right).$$

Combining this with step 2 proves the result. ■

Theorem 7 (Neighborhood robustification)**.** *Under assumptions 5 and 6, Algorithm 5 is a PAC-active auditor For $p>0$, a sample complexity of $N_0(\varepsilon,\delta/2)\left\lceil\frac{2}{(1-2\beta)^2}\log\frac{2N_0(\varepsilon,\delta/2)}{\delta}\right\rceil$ suffices in the presence of fairness-aware adversarial model owner.*

If the neighborhood samples are obtained by rejection sampling from an i.i.d unlabeled stream, the expected number of unlabeled draws is at most $\left\lceil\frac{2N_0(\epsilon,\delta/2)}{\gamma(1-2\beta)^2}\log\frac{2N_0(\varepsilon,\delta/2)}{\delta}\right\rceil$

Proof Fix $p>0$ and set $N=N_0(\epsilon,\delta/2)$.

Consider the t-th CGQ requested by k-ALEBI, conditionned on all previous CGQ responses being uncorrupted. By Lemma 19 and the definition of R,

$$\mathbb{P}\Big(\widehat{Y}_t\neq f^\star(q_t)\,\Big|\,\text{all previous CGQs are clean}\Big)\leq\frac{\delta}{2N}$$

There are at most N requested CGQs. Therefore

$$\mathbb{P}(\text{at least one CGQ response is corrupted}) \leq \frac{\delta}{2}$$

When running ROBUST k-ALEBI and k-ALEBI with the same unlabeled sample and the same internal randomization, on the event that every CGQ response is correct, and both executions receive the same interaction history with $\mathbf{M}$ and therefore return the same output.

Let F_0 be the failure event of this clean execution. By assumption, we have $\mathbb{P}(F_0) \leq \frac{\delta}{2}$. Hence the failure event of the robust algorithm is contained in $\{\text{there exists CGQ response that was corrupted}\} \cup F_0$. A union bound proves the (ϵ, δ)-PAC guarantee of learning the probe.

Finally, by definition of N, at most N clean CGQ requests are sent to $\mathbf{M}$, each using exactly R oracle attacks. This proves the upper bounds sample complexity. Each accepted local CGQ requires at most $1/\gamma$ unlabeled draws in expectation, which gives a sample complexity of $\left\lceil \frac{2N_0(\epsilon,\delta/2)}{\gamma(1-2\beta)^2} \log \frac{2N_0(\varepsilon,\delta/2)}{\delta} \right\rceil$.

■

Theorem 18 (Fairness lower bound under probabilistic attack)**.** *Under assumption 17 and* $0 < p < 1/2$*. A set of CGQs of size* $\Omega\big(\min\big\{\mathfrak{s}_\mu(\mathcal{F}), \frac{1}{\varepsilon}\big\}\big)$ *is necessary to learn the undelrying probe.*

Proof

We first prove the lower bound when the $\mathbf{M}$ is honest.

We fix a fairness-effective star of size m with center f_0, witnesses $f_1, \ldots, f_m$, and support $q_1, \ldots, q_m$, and let $\nu = \frac{1}{m}\sum_{j=1}^{m} \delta_{q_j}$.

Since f_0 is zero on the entire support, $\mu(f_0) = 0$. For every i, f_i is zero except at q_i. Since $f_i(q_i) \neq \mathbf{0}$ and individual coordinates belong to $\{-1, 0, 1\}$, at least one coordinate has one in abs. Thus $\mu(f_i) = \frac{1}{m}$. By the choice of m, $\frac{1}{m} \geq 4\varepsilon$.

Now fix i and couple executions under f_0 and f_i using the same unlabeled sample and the same internal randomness. Until q_i is queried, the two histories are identical. Let A_i be the event that q_i is never queried. On A_i, both executions return the same estimate. But $\frac{1}{m} \geq 4\varepsilon$, thus ϵ-accuracy intervals around 0 and $1/m$ are disjoint. Therefore the common estimate cannot be correct under both targets.

Since each execution fails with probability at most δ, $\nu_{f_0}(A_i) \leq 2\delta$. Thus

$$\nu_{f_0}(q_i \text{ is queried}) \geq 1 - 2\delta$$

Summing over queries, $\mathbb{E}_{f_0}\left[|\{\text{distinct queried support points}\}|\right] \geq m(1-2\delta)$. The number of distinct queried CGQs is at most the total number of oracle requests, proving the desired result.

To prove the resulting lower bound in the fairness aware adversarial regime, one can observe that the first term follows from the same indistinguishability argument as in Theorem **??**. We prove the attack-dependent term.

Step 1: construct two attacked environments. Fix $i \in [m]$. Under P_0, the target is f_0. On J_i, the clean response is $\mathbf{0}$, and the attack returns $f_i(q_i)$ with probability p.

Under P_i, the target is f_i. On J_i, the clean response is $f_i(q_i)$, and the attack returns $\mathbf{0}$ with probability p. Outside J_i, use the same observation law under both environments.

After identifying the two possible responses on J_i with $\{0, 1\}$, one observation from C_i has law

$$\mathrm{Bern}(p) \quad \text{under } P_0, \qquad \mathrm{Bern}(1-p) \quad \text{under } P_i.$$

Step 2: compute the history divergence. Let N_i be the number of oracle requests in J_i. Since the two environments differ only there, the chain rule for adaptive KL divergence gives

$$D_{\mathrm{KL}}(P_0 \| P_i) = \mathbb{E}_0[N_i] D_{\mathrm{KL}}\left(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(1-p)\right) \tag{15}$$

Step 3: Distinguish the environments. As proved previously, $\mu(f_0) = 0$, and $\mu(f_i) = \frac{1}{m} \geq 4\varepsilon$. Thresholding $\widehat{\mu}$ at $1/(2m)$ gives a binary test between P_0 and P_i with both error probabilities at most δ. By data processing,

$$D_{\mathrm{KL}}(P_0 \| P_i) \geq D_{\mathrm{KL}}\left(\mathrm{Bern}(\delta) \,\|\, \mathrm{Bern}(1-\delta)\right)$$

With 15 gives

$$\mathbb{E}_0[N_i] \geq \frac{D_{\mathrm{KL}}\left(\mathrm{Bern}(\delta) \,\|\, \mathrm{Bern}(1-\delta)\right)}{D_{\mathrm{KL}}\left(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(1-p)\right)}.$$

Summing over cells for $i = 1, \ldots, m$,

$$\mathbb{E}_0[T] \geq m \frac{D_{\mathrm{KL}}\left(\mathrm{Bern}(\delta) \,\|\, \mathrm{Bern}(1-\delta)\right)}{D_{\mathrm{KL}}\left(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(1-p)\right)}.$$

Combining this with the indistinguishability lower bound proves

$$\mathbb{E}[T] \geq m \max\left\{1 - 2\delta, \frac{D_{\mathrm{KL}}\left(\mathrm{Bern}(\delta) \,\|\, \mathrm{Bern}(1-\delta)\right)}{D_{\mathrm{KL}}\left(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(1-p)\right)}\right\}$$

$$\mathbb{E}[T] \geq m \max\left\{1 - 2\delta, \frac{(1-2\delta)\log\frac{1-\delta}{\delta}}{(1-2p)\log\frac{1-p}{p}}\right\}$$

. ■

As the corruption increases and p gets closer to $1/2$,

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(1-p)\right) = 2(1-2p)^2 + O((1-2p)^4),$$

while for δ close to O

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(\delta) \,\|\, \mathrm{Bern}(1-\delta)\right) = \Theta(\log(1/\delta))$$

THus, in the high-noise regime and high-confidence regime, the sample complexity becomes comparable to the one derived in Theorem 14 $\Omega\Big(\min\left\{\mathfrak{s}_\mu(\mathcal{F}), \frac{1}{\varepsilon}\right\} \frac{\log(1/\delta)}{(1-2p)^2}\Big)$

G Auxiliary Definitions and Lemmas

Theorem 20 (Distirbution-free realizable multiclass learning **?**). *Let $\mathcal{F}_k$ be a multiclass hypothesis class of Daniely-Shwartz dimension DS_k. Then there exists a learning algorithm $\mathcal{A}$ such that, a sample complexity of $\left(\frac{\mathrm{DS}_k+\log(1/\delta)}{\epsilon}\right)$ is sufficient to learn $\mathcal{F}_k$ in the distirbution-free setting under realizability.*

Definition 10 (Extended Teaching Dimension). *Let $\mathcal{V} \subseteq \mathcal{Y}^m$ be a finite hypothesis class over the finite domain $[m] = \{1, \ldots, m\}$. For an arbitrary labeling $g \in \mathcal{Y}^m$, a set $S \subseteq [m]$ is said to specify g with respect to $\mathcal{V}$ when*

$$|\{v \in \mathcal{V} : v|_S = g|_S\}| \leq 1.$$

The extended teaching dimension of V is

$$\mathrm{XTD}(\mathcal{V}) \triangleq \max_{g \in \mathcal{Y}^m} \min \{|S| : S \subseteq [m] \text{ specifies } g \text{ with respect to } \mathcal{V}\}.$$

Lemma 21 (Freedman's inequality Freedman (1975); Tropp (2011)). *Suppose that $D_t \leq b$ almost surely for every t. Then, for every $s \geq 0$ and $v > 0$,*

$$\mathbb{P}\left(S_T \geq s,\ V_T \leq v\right) \leq \exp\left(-\frac{s^2}{2(v + bs/3)}\right).$$

More generally, the original Freedman inequality is uniform over time.

H Additional Experiments Results and Details

H.1 Interpretability

In addition to estimating multi-group fairness, k-ALEBI provides an interpretation of the measured unfairness that can't be obtained from a scalar fairness estimate alone. Its interpretability outputs decompose the learned probe across non-protected features and characterize how different feature configurations affect the CGQs response. For a feature q, let $Z_q(Q) = (X_1^{(q)}, \ldots, X_K^{(q)})$ and recall $\mathfrak{e}_{q,ij}(z) = \mathbb{E}[f_{ij}(Q) \mid Z_q(Q) = z]$, $\widetilde{f}_q(z) = \max_{i<j} |m_{q,ij}(z)|$. We estimate these conditional means using independent i.i.d CGQs and K-nearest-neighbor regression at the observed Z_q configurations. For $K = 2$, we additionally visualize $\widetilde{f}_q(z_1, z_2)$ using Gaussian-kernel surfaces. Importantly, k-ALEBI observes CGQs responses rather than individual model predictions.

Figure 4 ranks features according to the mean across used seeds of the 90th percentile of the estimated $\widetilde{f}_q(Z_q)$. The corresponding feature-pair visualizations retain information about the direction and configuration of group disparities with respect to $\mathbf{M}$ that would otherwise be lost by reducing them to a single unsigned score. Thus the analysis identifies not only the features associated with larger measured disparities, but also the group pairs and feature configurations under which the learned model exhibits different CGQd responses. For the COMPAS dataset, Figure 5 indicates larger estimated disparities for individuals of non-Caucasian race who had been arrested multiple times, with the disparity varying across the number of prior arrests ($n \in \{1, 2, 5\}$). For the Student dataset, past-class failures and the amount of time spent going out are associated with larger gender-related disparities, with the estimated disparity being larger for females than for males. The feature "father's education" also exhibits a disparity pattern: higher levels of paternal education are associated with larger estimated differences in the model's group disparities for males.

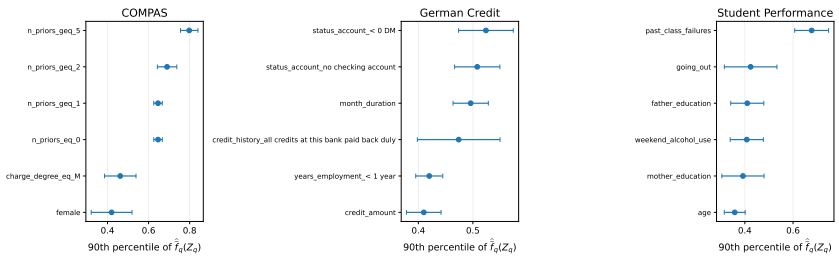

Figure 4: Feature-wise ranking of probe-derived disparity scores.

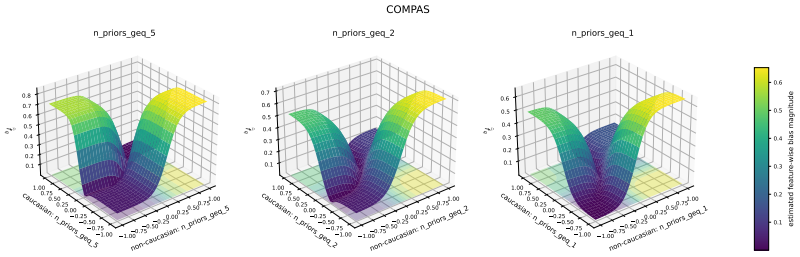

Figure 5: Feature-wise visualization of learned probes on the COMPAS dataset.

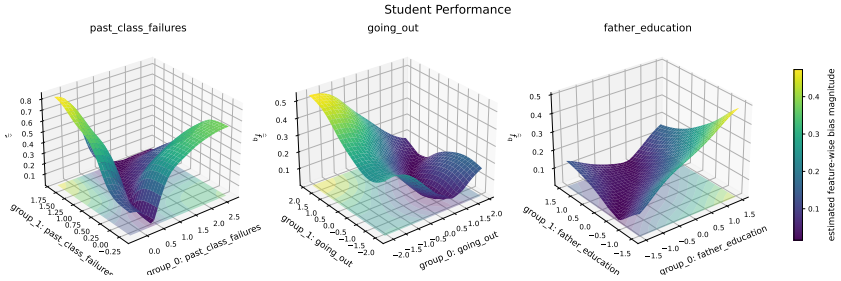

Figure 6: Feature-wise visualization of learned probes on the Student dataset.

H.2 Protection Against Model Extraction Attacks

We next examine the distinction between accurately auditing fairness and revealing sufficient information to reconstruct the underlying model. At limited query budgets, conditional group queries (CGQs) can preserve a large version space of models with similar discriminative behavior while still providing accurate fairness estimates. This demonstrates an empirical separation between audit utility and immediate model identification: useful fairness information can be obtained without necessarily exposing the individual predictions required for direct model reconstruction.

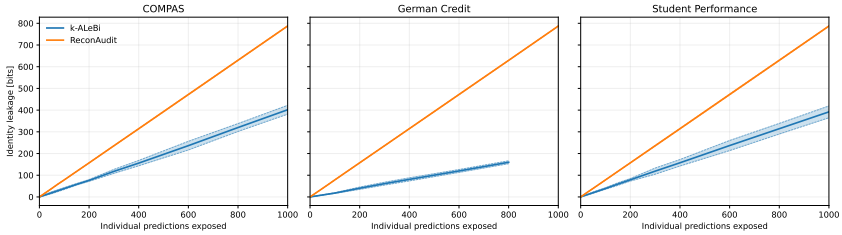

Figure 7: Identity leakage, measured by the number of individual model responses revealed to the auditor, comparing k-ALeBi with ReconAudit.

Figure 7 compares the identity leakage induced by k-ALeBi and ReconAudit. ReconAudit uses membership queries by default, which directly expose individual model labels to the auditor. In contrast, k-ALeBi operates through relational CGQs and therefore does not directly reveal the corresponding individual labels. We quantify this information exposure using the identity-leakage measure defined in Definition 8. Across the evaluated query budgets, the results show substantially lower identity leakage for k-ALeBi than for ReconAudit, while retaining the ability to perform fairness auditing.

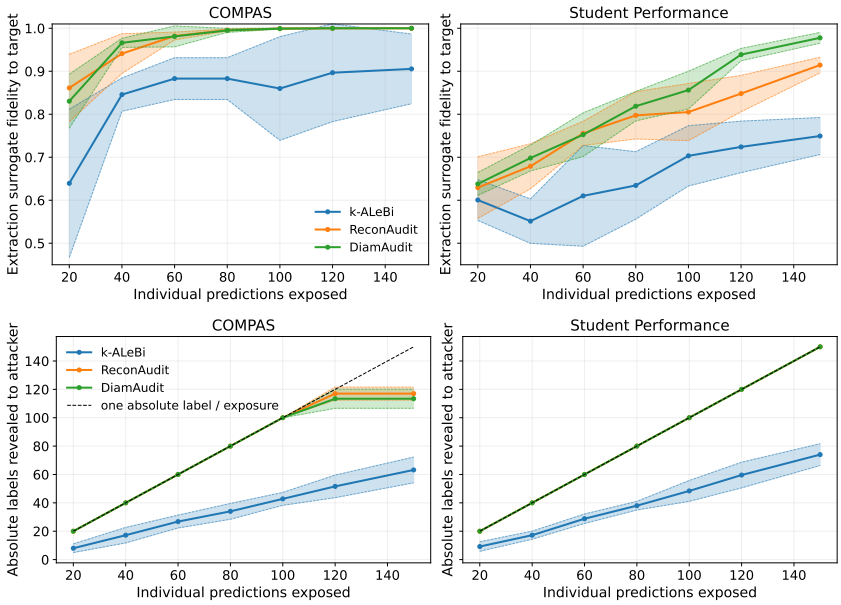

Figure 8: Model-extraction behavior under k-ALEBI and DIAMAUDIT: extraction surrogate fidelity to the target model and the number of absolute labels revealed to the attacker.

Figure 8 further compares k-ALEBI with DIAMAUDIT. The first panel measures the fidelity of an extraction surrogate to the target model as individual predictions are exposed, while the second reports the number of absolute labels revealed to the attacker. These results illustrate the information advantage of relational auditing: k-ALEBI can provide useful fairness information through CGQs without requiring the same level of exposure of individual model predictions as membership-query-based extraction.

Overall, the results indicate that CGQs can reduce the amount of model-identifying information exposed during auditing. This allows fairness properties to be investigated while limiting direct access to individual model responses, thereby separating the utility of fairness auditing from the immediate reconstruction of the underlying model.

H.3 ROBUSTNESS AGAINST FAIRNESS-AWARE ADVERSARIES

Finally, we evaluate the robustness of ROBUST k-ALEBI against a fairness-aware adversary. On COMPAS-linear, for example, under a high corruption regime of $p = 0.40$, the adversary induces an average of 41.4 raw coordinate corruptions, of which only 6.2 result in corrupted queries after filtering. Despite these corruptions, the fairness estimation error increases only marginally, from 0.019 to 0.021, with no observed concealment across the five runs. In comparison, the reconstruction error of RECONAUDIT increases from 0, corresponding to exact model identification in the absence of corruption, to 0.074 under attack, with a concealment rate of 0.20.

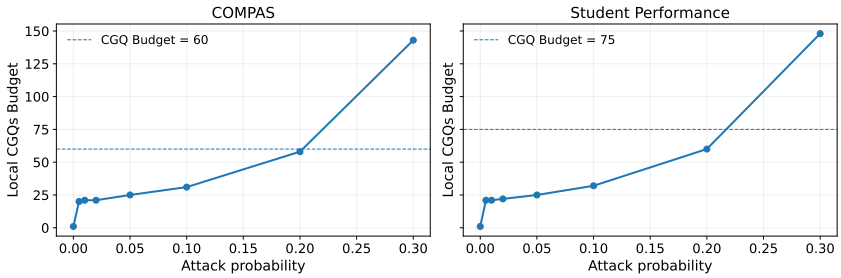

Figure 9: Local CGQ budget of ROBUST k-ALEBI across cells as a function of the attack probability.

The Yan & Zhang (2022) baseline exhibits an estimation error of 0.03 in the absence of attack, which increases to 0.05 under the strongest feasible attack probability, capped at 0.12. The same qualitative stability of ROBUST k-ALEBI is observed on the German-linear and Student-linear models.

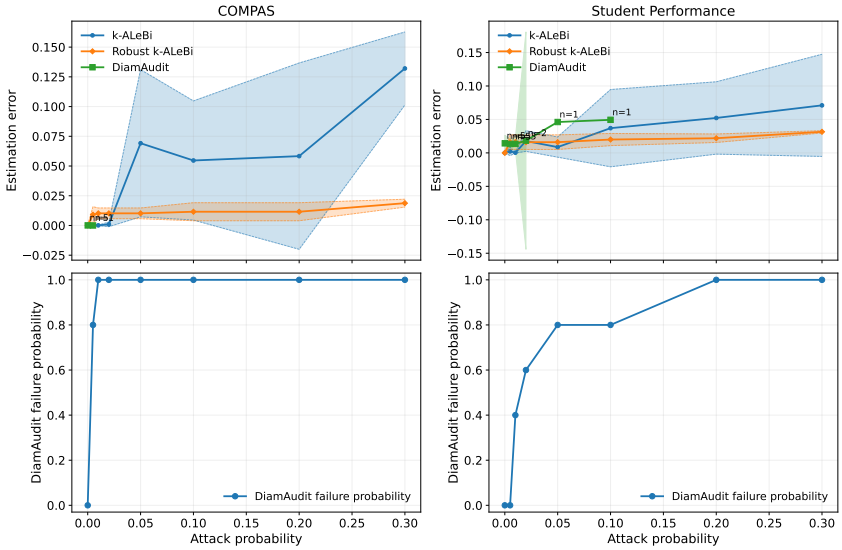

Figure 10: Estimation error for statistical parity as a function of the attack probability.

H.4 FAIRNESS-AWARE ADVERSARIAL REGIME

The finite-pool implementation follows the neighborhood mechanism above, but constructs the cells using a concrete target-independent pseudometric.

Let G be a fixed reference panel of candidate probe hypotheses, selected without knowledge of the deployed target. Define

$$d_G(q, q') := \frac{1}{|G|} \sum_{f \in G} \mathbf{1}\{f(q) \neq f(q')\}$$

For every candidate CGQ q, its cell $J(q)$ is fixed before any oracle interaction and consists of the 256 nearest distinct CGQs under d_G. Neither the current version space nor observed target responses are used to construct these cells.

For $p > 0$, the implementation in the experimental section uses

$$\beta = p + \rho - p\rho, \qquad \rho = 0.05, \qquad \delta = 0.05,$$

and

$$R(p) = \left\lceil \frac{2}{(1-2\beta)^2} \log \frac{2N}{\delta} \right\rceil,$$

where N bounds the number of clean CGQ decisions allowed by budget?

When ROBUST k-ALEBI requests q_t, it also samples $R(p)$ distinct CGQs from $J(q_t)$ without replacement and receives response $\widetilde{Y}_{t,1}, \ldots, \widetilde{Y}_{t,R}$. Then , $\widehat{Y}_t = \text{mode}\{\widetilde{Y}_{t,1}, \ldots, \widetilde{Y}_{t,R}\}$, and uses this as the estimated response of q_t.

If the remaining exposure budget cannot pay for all $R(p)$ local queries, ROBUST k-ALEBI abstains rather than using a smaller block. When $p = 0$, the cell size is consummed, so ROBUST k-ALEBI coincides with k-ALEBI.

The implemented attacks use independent Bernoulli-p attack, which are a special case of Assumption 6. When an attack occurs, the experiment chooses an incorrect relational response that increases the current absolute statistical-parity estimation error.

H.5 AUDITORS BASELINES

The supplied Yan & Zhang (2022) implementation DIAMAUDIT performs a search in the current linear version space for hypotheses with high and low statistical-parity values and queries a point on which they disagree. The Yan curves use the original version-space average statistical-parity error computed by a hit-and-run evaluation protocol on the model class. For k-ALEBI, we use a Monte-Carlo approximation of the same linear-model family. A CGQ contains one individual from each group and therefore may cost up to two exposed individual predictions per coordinate $i \leq k$. The horizontal axis measures these individual predictions exposed allowing model extraction attacks.

The original implementation of Yan & Zhang (2022) baseline is not run on German Credit. The supplied implementation assumes a binary protected attribute, whereas our German Credit setup contains four `status_and_sex` groups, for a total of $k = 6$ groups.

For the plotted comparison where $k = 2$, k-ALEBI and DIAMAUDIT are evaluated using the same statistical-parity unfairness $\mu = |P(h = 1 \mid A = 1) - P(h = 1 \mid A = 0)|$ with absolute value estimation error $|\widehat{\mu} - \mu^\star|$. The original saved Yan & Zhang (2022) endpoint appears only in the validation table because it retains the evaluation convention of their implementation.

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.