On the Hardness of Auditing Model Properties Under Updates: Complexity of Property-Preserving Updates

Ayoub Ajarra **Debabrota Basu**
Équipe Scool, Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189-CRIStAL, Lille, France

Abstract

As Machine Learning (ML) becomes deeply embedded in societal infrastructure, assessing the risks posed by these models has grown increasingly critical. Real-world deployment further complicates this assessment: model owners may apply strategic updates in response to dynamic environments (e.g., financial markets), potentially undermining key guarantees. We formalize this setting and address two goals: (i) *accurately estimating a target auditing property– such as group fairness– using a minimal number of labeled samples*; and (ii) *characterizing the complexity of strategic updates by identifying the subset of admissible updates that preserve the property*. To this end, we propose a generic algorithmic framework for efficient PAC auditing, powered by an Empirical Property Optimization (EPO) oracle. For statistical parity, we establish distribution-free audit bounds characterized by the SP dimension, a new combinatorial measure that captures the complexity of admissible strategic updates. Finally, we show that our framework naturally extends to other properties, including prediction error and robust risk. Code is available at `https://github.com/AyoubAjarra/Auditors-with-prospects`.

1 INTRODUCTION

The rise of algorithmic decision makers in high-stakes domains such as healthcare, finance, and employment (Aboy et al. 2024; Montag and Finck 2024; Obermeyer et al. 2019; Raji and Buolamwini 2019) has made algorithmic regulation a central challenge. While regulation can take different forms (Vecchione et al. 2021), accurately auditing properties of stochastic ML models has emerged as the computational bedrock. In trustworthy ML, the focus is to ensure that ML models meet social and ethical constraints, such as fairness measures, robustness, and privacy guarantees, often through accurate assessment of property violations (also known as, *auditing*) (Le Merrer et al. 2023; Madaio et al. 2020; Raji et al. 2020). This has spurred significant interest in ML communities to develop accurate and reliable auditing algorithms (Ajarra et al. 2024; Cohen et al. 2019; Hsu et al. 2024; John et al. 2020; Kearns et al. 2018; Neiswanger et al. 2021; She et al. 2025; Yan and Zhang 2022).

However, real-world assessments of algorithmic bias are complex due to model drift and can fundamentally alter the very properties under audit (Lu et al. 2018; Widmer and Kubat 1996). These shifts may occur in several scenarios. For instance, in evolving operational scales, a business (managed by a model owner) expands and gains access to vastly larger and more diverse datasets (e.g., customer transactions and market signals). Under such conditions, simple models like linear classifiers or decision trees may underfit, failing to capture the emergent complexity of data. In particular, the underlying model class itself may change. For example, a model trained under stable market conditions using linear assumptions may perform poorly when markets become highly volatile, as nonlinear dependencies emerge (Cont 2001). In such regimes, the model owner needs to update its model to flexible architectures, such as neural networks, that potentially yield superior predictive performance (Goodfellow et al. 2016). This motivates new auditing frameworks that allow model updates without compromising the audited property. Yan and Zhang (2022) studies the auditing of Statistical Parity (Definition 1) *under the assumption that the auditor knows the hypothesis class $\mathcal{F}$ to which the model under audit belongs*. They introduce manipulation-proofness as a robustness criterion for auditing procedures, which is achieved by

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

constructing a version space $\mathcal{V}(S) \subseteq \mathcal{F}$ induced by a teaching set S. Specifically, the version space is defined as $\mathcal{V}(S) = \{f' \in \mathcal{F} : f'(x) = f(x),\ \forall (x, y) \in S\}$, where $f \in \mathcal{F}$ is the model under audit. An audit is said to be manipulation-proof if the auditor's estimate of the fairness property remains invariant for all models in $\mathcal{V}(S)$. In other words, once the auditor constructs $\mathcal{V}(S)$ based on the observed input-output S, any strategic update to the model that remains within $\mathcal{V}(S)$ cannot change the audit outcome. Crucially, the size of this teaching set scales with the sample complexity required to *reconstruct the model*. Consequently, if the model owner is constrained from altering predictions on these audit points, the model itself remains essentially unchanged with high probability. This effectively precludes meaningful model updates during the audit process, rendering the manipulation-proofness condition overly restrictive in general applications. Moreover, their approach assumes a known hypothesis class $\mathcal{F}$ (i.e., linear classifiers) and does not account for model updates that alter the underlying model class itself, a scenario that is common in practice when models are retrained, fine-tuned, or replaced with architectures of different structure. In this work, we extend the audit setting by allowing arbitrary updates to the post-audit model from a (potentially) different model class $\mathcal{F}$. We refer to such updates not altering the property under audit as prospects. On the other hand, Yan and Zhang (2022) focuses on reconstructing the model using a fixed model class, and thus, characterizes the complexity of auditing under manipulation-proofness via classical learning-theoretic measures (e.g., the disagreement coefficient and VC-dimension). They leave open the question of whether alternative information-theoretic measures could yield a more flexible characterization of audit complexity. In this work, we resolve this question in the non-interactive regime, in which the auditor has a one-shot access to samples. We ask:

Q1: *Given a strategic class $\mathcal{F}$, what is the information complexity of auditing group fairness without full model reconstruction* [1] *?*

Q1 seeks a combinatorial characterization of the classes of strategic updates $\mathcal{F}$ for which group fairness auditing with prospects is feasible. In particular, it asks how the complexity of $\mathcal{F}$ governs the amount of information required to certify fairness properties without learning the model itself. As model classes grow increasingly expressive — most notably with the use of neural networks that go far beyond linear decision boundaries — a complementary question naturally arises:

[1]For example, in terms of learning-theoretic complexity measures such as VC dimension

Q2: *If the strategic class grows arbitrarily in VC dimension, can strategic updates still admit information-theoretically feasible fairness auditing?*

In highly overparameterized regimes, learning is known to become information-theoretically hard, with VC dimension serving as a proxy for combinatorial complexity. **Q2** investigates whether fairness auditing may nevertheless remain strictly easier than learning, even for highly expressive model classes.

Related work. There are two main lines of work in this context. One that tries to *verify* whether a property of an ML model shoots over a certain threshold. This has been extensively studies in the case of robustness (Cohen et al. 2019; Salman et al. 2019), group fairness measures, like SP (Albarghouthi et al. 2017; Neiswanger et al. 2021), and individual fairness (John et al. 2020). Hsu et al. (2024); Kearns et al. (2018) aimed to verify (SP) through a reduction to weak agnostic learning. Hsu et al. (2024) studied this framework for Gaussian feature distributions and homogeneous halfspace subgroups, and demonstrate the problem's computational hardness. Though the initial literature focused on verification, it requires a priori knowledge of a valid threshold, which is hard to pre-define without social and application context. The other line of works try to accurately and statistically *estimate* the property under audit. Neiswanger et al. (2021) proposes a Bayesian approach to estimate distributional properties. Wang et al. (2022) estimates simpler distributional properties, e.g. mean, median. For ML models, Yan and Zhang (2022) further propose active learning algorithms to sample efficiently and audit SP of a model under the assumption that its class is known *a priori*. A detailed discussion on related work is deferred to Appendix A.

Contributions: A New Framework. Our work introduces a general framework for the auditing problem under strategic updates that applies to *unknown* strategic classes. We shift the goal of the auditor from merely estimating the property value to also requiring the auditor to output a subset of the strategic class that preserves this property under audit (*prospect class*, Definition 6). This generalizes the standard notion of manipulation-proofness to settings involving wider range of model updates.

1. **Universal Auditor.** We propose a generic algorithm EPO (Algorithm 1) framework that, given the property of interest, outputs an estimate of the property and the prospect class. Specifically, we define auditing losses for different properties: SP, learning error, learning stability, and robust risk.

2. **Fairness audits and SP dimension**: We focus on statistical parity as our primary property of

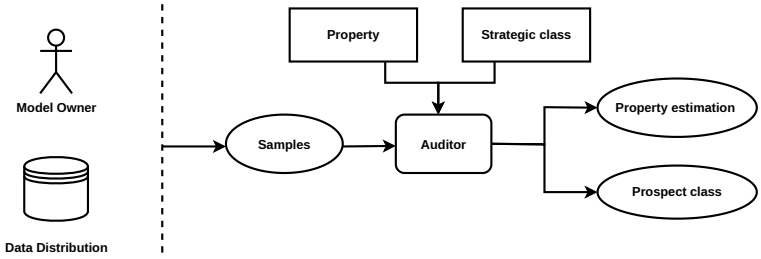


Figure 1: A schematic of black-box auditing with prospects.

interest. We provide sample complexity guarantees for finite and infinite hypothesis classes. We introduce a new capacity measure called the SP dimension to weakly audit infinite hypothesis classes (Definition 2). We demonstrate its relationship to VC dimension. For multiple protected groups, we prove that any shattered set must include instances from distinct protected groups, providing new insights into multi-group auditing scenarios.

3. **Measuring coverage of prospect class**: We introduce the prospect ratio as a data-dependent measure to measure the coverage of the prospect class. For statistical parity, we provide concentration bounds on the estimation of the prospect ratio.

Finally, we numerically demonstrate that our framework yields good estimates of SP and prospect class for real-life datasets and multiple ML models.

2 PRELIMINARIES

In this section, we present the problem formulation and introduce illustrative examples of properties of ML models, leading to a formalization of the PAC auditing framework under model updates in two distinct settings (i.e., weak and strong auditing).

Notations. Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output spaces of an ML model, respectively. The input data follows a distribution $\mathcal{D}_{\mathcal{X}}$. $\mathcal{D}$ denotes the joint distribution over the product space $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{F}$ denotes a class of models from $\mathcal{X}$ to $\mathcal{Y}$. An unknown and possibly randomized learning model $f : \mathcal{X} \to \mathcal{Y}$ labels the inputs from $\mathcal{X}$. A property of the ML model is defined as a functional $\mu : \mathcal{F} \times \mathcal{P} \to \mathbb{R}$ that takes the model and the corresponding data-generating distribution to yield a real number, where $\mathcal{P}$ is a class of generating distributions.

2.1 Examples of properties

Statistical parity measures discriminative bias in positive predictions between two protected groups (Feldman et al. 2015).

Definition 1 (Statistical Parity). *Let $\{\mathcal{X}_0, \mathcal{X}_1\}$ denote a partition of the input space based on a binary protected attribute (e.g., gender, where $\mathcal{X}_0$ corresponds to females and $\mathcal{X}_1$ to males). The statistical parity of a model $f \in \mathcal{F}$ measures the discrepancy of its predictions with respect to the two protected groups $\mathcal{X}_0$ and $\mathcal{X}_1$, i.e.* $\mu(f, \mathcal{D}) \triangleq |\underset{x\sim\mathcal{D}_{\mathcal{X}}}{\mathbb{P}}[f(x) = 1|x \in \mathcal{X}_0] - \underset{x\sim\mathcal{D}_{|\mathcal{X}}}{\mathbb{P}}[f(x) = 1|x \in \mathcal{X}_1]|$.

A model is fair across protected groups if it achieves small statistical parity value. Definition 1 can be extended to multiple protected groups by the maximum discrepancy between groups.

Expected risk. Measures the expectation of prediction errors by a model f on input-output pairs generated by $\mathcal{D}$ (Mohri et al. 2018). For the class of binary classifiers, $\mu(f, \mathcal{D}) = \underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}[f(x) \neq y]$.

Learning stability. Measures the discrepancy in predictive performance across two environments. For a model trained on a distribution $\mathcal{D}_{\text{src}}$ and deployed on $\mathcal{D}_{shift}$. We define its learning stability as $\mu(f, \mathcal{D}_{\text{src}}, \mathcal{D}_{\text{shift}}) \triangleq \left| \underset{(x,y)\sim\mathcal{D}_{\text{src}}}{\mathbb{P}}[f(x) \neq y] - \underset{(x,y)\sim\mathcal{D}_{\text{shift}}}{\mathbb{P}}[f(x) \neq y] \right|$. This definition is closely related to distribution shifts; however, distribution shift itself is a property of the data alone: $\mathcal{D}_{\text{src}} \neq \mathcal{D}_{\text{shift}}$ (Ben-David et al. 2010; Taori et al. 2020). Our definition evaluates how stable the model's performance remains under such shifts. Details are given in Appendix B.

Robust Risk. After deploying an ML model, we come across inputs that are noisy or manipulated by an adversary. To ensure safety, it is necessary to verify that the ML model is robust against such perturbations of the input. Robustness is measured using robust risk and is further ensured by minimizing the robust risk (e.g., works on adversarial ML (Liu et al. 2021)). For any x in $\mathcal{X}$, let $\mathcal{U}(x)$ denote the set of perturbations acting on input x, and let $\mathcal{D}^*$ denote the resulting distribution after the true model deployment. Robust risk is defined as $\mu(f, \mathcal{D}, \mathcal{U}) = \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{1}_{f(z)\neq y}\right]$.

2.2 Auditing Problems

As discussed in the introduction and illustrated in Figure 1, the auditor is given access to a strategic class $\mathcal{F}$, takes as input a set of i.i.d samples (labeled by the post-audit model and inaccessible to the auditor) and the desired property μ, and outputs both an estimate of the property and a property-preserving class that we refer to as the prospect class. Formally, an auditing problem is defined as a quintuplet $\langle \mathcal{X}, \mathcal{Y}, \mathcal{F}, \mathcal{P}, \mu, \ell \rangle$, where f is the model under audit. ℓ is a loss that depends *implicitly* on the property under audit μ. Details on how the loss function depends on the audited property are provided in Appendix B, where we define loss functions for the properties discussed above and show how our framework extends to broader auditing problems. In the following, we make no assumptions about f. In particular, we do not assume knowledge of the model class. Indeed, f may be a randomized machine learning algorithm whose behavior depends on an unknown randomization mechanism. In the next section, we introduce a novel auditing setting that not only evaluates existing models but also yields prospective machine learning models for future deployment (aka Prospect class, see Definition 6). We establish a formal connection between auditing properties of black-box models with prospective guarantees and agnostic learning of those properties from data within a strategic class $\mathcal{F}$.

3 AUDITING WITH PROSPECTS

During the auditing process, the model owner provides the auditor with a finite set of samples labeled by a black-box model. Subsequently, the owner may wish to update their decision rule; this strategic update is communicated to the auditor. To accommodate such strategic updates, the auditor identifies the prospect class: subclass of the strategic class that preserves the audited property. Agnostic auditing refers to the scenario where the pre-audit model is not an element of the strategic class. The prospect class must include models that have the same property as the pre-audit model (defined as $\mathrm{OPT}(\mu, \mathcal{D}, \mathcal{F}) = \min_{f \in \mathcal{F}} |\mu_{\mathcal{D}}(f) - \mu(\mathcal{D})|$). Realizable auditing corresponds to the scenario where the pre-audit model belongs to the strategic class, and in this case $\mathrm{OPT}(\mu, \mathcal{D}, \mathcal{F}) = 0$. This motivates Definition 2, where in weak auditability, finding one model in the prospect class suffices. Finally, let $\mathrm{OPT}(\mu, S, \mathcal{F}) = \min_{f \in \mathcal{F}} |\mu_{S_{\mathcal{X}}}(f) - \mu(S)|$.

Definition 2 (Weakly μ-auditable class)**.** *We call an algorithm $\mathcal{A}$ to be (ϵ, δ)-weak auditor for a given auditing problem $\langle \mathcal{X}, \mathcal{Y}, h, \mathcal{F}, \mathcal{P}, \mu \rangle$ when it uses a sample set S of size m sampled from any $\mathcal{D} \in \mathcal{P}$ to yield a model $\mathcal{A}[S]$ and an estimate $\mu_{\mathcal{D}}(\mathcal{A}[S])$ satisfying:*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \Big[|\mu_{\mathcal{D}}(\mathcal{A}[S]) - \mathrm{OPT}(\mu, \mathcal{D}, \mathcal{F})| \geq \epsilon \Big] \leq \delta \,,$$

where $\epsilon, \delta \in (0,1)$ and m is a bounded function of the problem parameters and $(1/\epsilon, 1/\delta)$. We say that a strategic class $\mathcal{F}$ is weakly μ-auditable if for all $(\epsilon, \delta) \in (0,1)^2$, there exists an (ϵ, δ)-weak auditor.

Algorithm 1 EPO Oracle

Require: Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, Strategic class $\mathcal{F}$, Property to audit μ

Ensure: Estimate of the property $\hat{\mu}$, prospective model $\hat{f}$

1: Define empirical risk for μ: $\mathcal{E}_m(f, \mu)$

2: Use an ERM oracle to solve the optimization problem:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathcal{E}_m(f, \mu)$$

3: **return** $\hat{f}$ and $\hat{\mu}_m(\hat{f})$

From ERM to Auditors: Auditing with loss functions. Algorithm 1 proposes a generic framework for using ERM oracles to weakly audit any property μ of an ML model f given a strategic class $\mathcal{F}$. Broadly, the intuition is that if we can define a loss function to audit each of the properties, we can use the samples collected from the black-box model f to learn a prospective model $\hat{f} \in \mathcal{F}$. One can use any off-the-shelf ERM solver, like SGD, Adam, ADMM, etc., in Line 2. The following definition extends weak auditability by saturating the strategic class and identifying all models within the prospect class that satisfy the audited property.

Definition 3 (Strongly μ-auditable class)**.** *We call an algorithm $\mathcal{A}$ to be (ϵ, δ)-strong auditor for a given auditing problem $\langle \mathcal{X}, \mathcal{Y}, \mathcal{F}, \mathcal{P}, \mu \rangle$ when it uses a sample set S of size m sampled from any $\mathcal{D} \in \mathcal{P}$ to yield a class of models $\mathcal{A}[S] \subseteq \mathcal{F}$ and an estimate $\mu_{\mathcal{D}}(\mathcal{A}[S])$ satisfying:*

(i) ***Correctness:***

$$\mathbb{P}_{S \sim \mathcal{D}^m} \Big[\sup_{f \in A[S]} |\mu_{\mathcal{D}}(f) - \mathrm{OPT}(\mu, \mathcal{D}, \mathcal{F})| \geq \epsilon \Big] \leq \delta$$

.

(ii) ***Completeness:***

$$\mathbb{P}_{S \sim \mathcal{D}^m} \Big[\inf_{f \in A^C[S]} |\mu_{\mathcal{D}}(f) - \mathrm{OPT}(\mu, \mathcal{D}, \mathcal{F})| \leq \epsilon \Big] \leq \delta$$

.

$A[S] \subseteq \mathcal{F}$ is a subclass of models that achieve a small error on the training set S. We say that a hypothesis class $\mathcal{F}$ is strongly μ-auditable if for all $(\epsilon, \delta) \in (0,1)^2$, there exists an (ϵ, δ)-strong auditor.

Strong auditability is governed by two key conditions: correctness, which requires all models in the prospect class to preserve the same property as the pre-audit model, and completeness, which ensures the hypothesis class has been fully exhausted by including all models that share the property μ with the pre-audit model. The resulting prospect class serves as the set of potential post-audit models.

4 AUDITING STATISTICAL PARITY

In this section, we focus on characterizing the problem of auditing statistical parity. In the black-box setting, the auditor can only access a pool of labeled instances by f, which we denote S. This is equivalent to assuming access to an empirical distribution $\hat{\mathcal{D}}_S$ that encodes the discriminative behavior of the black box model with respect to the fixed protected groups. From now on, we denote the protected groups by $\mathcal{X}_0$ and $\mathcal{X}_1$, which form a partition of the input space $\mathcal{X}$. Similarly, for a finite sample S, let $S_0 \subseteq \mathcal{X}_0$ denote samples from the first protected group and $S_1 \subseteq \mathcal{X}_1$ from the second protected group.

4.1 Weakly Auditable classes

We characterize the complexity using the minimum number of samples required for each protected group (m_0 and m_1). This approach differs from existing methods that rely on the total sample size ($m = m_0 + m_1$), which requires assuming specific proportions between protected groups Yan and Zhang (2022). By considering the minimum samples needed for each group independently, our method remains valid regardless of how the probability mass is concentrated across protected groups, making it more robust to group imbalances in real-world scenarios. To rigorously define PAC weakly auditing of statistical parity, we denote by $\mathcal{F} \subseteq 2^{\mathcal{X}}$ the strategic class (a generalization of the hypothesis class). With this structure, we present the strategic lemma linking EPO oracle solvers over pairs of points from different protected groups to uniform convergence in statistical parity.

Lemma 1 (Strategic Lemma). *Let $\epsilon, \delta \in (0,1)$, $m : (0,1)^2 \to \mathbb{N}$. Suppose that the following holds:*

- ***Estimation accuracy.*** *$\mathcal{A}$ outputs f_S from $\mathcal{F}$: $\mathbb{P}_{S\sim\mathcal{D}^m}\Big[|\hat{\mu}_S(f_S) - \widehat{OPT}(S,\mathcal{F})| > \frac{\epsilon}{3}\Big] < \frac{\delta}{2}$.*
- ***Uniform convergence.*** *$\mathcal{F}$ verifies $\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\exists f \in \mathcal{F}, |\mu_{\mathcal{D}}(f) - \hat{\mu}_S(f)| > \frac{\epsilon}{3}\Big] < \frac{\delta}{2}$.*

Then, $\mathcal{A}$ is (ϵ,δ)-weak auditor for statistical parity.

Lemma 1 establishes that uniform convergence over the property, combined with empirical audit risk minimization, guarantees weak auditing PAC auditing of that property. This lemma is risk-free, unlike the one used to derive VC bounds, which uses the intermediate of a loss function that we do not consider here. The proof is given in Appendix C.

Finite Strategic Class. We first present a result on the sample complexity for weakly auditable finite classes.

Theorem 2 (Agnostic weak auditability). *If $\mathcal{F}$ is a finite hypothesis class, than $\mathcal{F}$ is weakly auditable, with respect to statistical parity, for any distribution on $\mathcal{X}\times\mathcal{Y}$ with a sample complexity $\mathcal{O}\Big(\Big\lceil \frac{18}{\epsilon^2}\log\frac{8|\mathcal{F}|}{\delta}\Big\rceil\Big)$.*

Theorem 2 offers an intuitive understanding of the auditing hardness, highlighting in the case of a finite class that auditing SP using weak auditing framework requires more sample complexity than both learning and reconstructing the model. The proof is given in Appendix D.1.

Infinite Hypothesis Class. The finiteness of $\mathcal{F}$ poses inherent limitations for auditing, as such a hypothesis class may not contain models that satisfy the desired post-audit stakes. This may result in a singleton prospect class (pre-audit model), rendering the auditing process ineffective. Therefore, we extend our analysis to infinite strategic classes. Here, VC dimension falls short in tightly measuring the complexity of SP auditing. We illustrate this in the following example:

Example 1. *Consider a set $\mathcal{X} \subseteq \mathbb{R}^2$ where the protected attribute is 'gender' and the second feature is 'age'. In the context of classification in $\mathbb{R}^2$ using linear classifiers (Figure 2), it is known that the VC dimension of this class is three. However, if the three points that can be shattered by the class of linear classifiers share the same protected attribute (gender), they are collinear and cannot be shattered.*

Next, we define group traces of the strategic class $\mathcal{F}$ with respect to the protected groups $\mathcal{X}_0$ and $\mathcal{X}_1$.

Definition 4 (Group-Traces of a strategic class). *Let $\mathcal{X}$ denote an uncountable space, $\mathcal{F}$ a set of subsets of $\mathcal{X}$, and S a finite subset of $\mathcal{X}$. The group-traces of $\mathcal{F}$ in the protected groups of $S = S_0 \cup S_1$, denoted by $\Delta^{SP}_{\mathcal{F}}(S)$, is defined as $\Delta^{SP}_{\mathcal{F}}(S_0,S_1) \triangleq \Big\{(A_0,A_1) | A_0 \subseteq S_0, A_1 \subseteq S_1, \exists c \in \mathcal{F}, A_0 = c \cap S_0, A_1 = c \cap S_1\Big\}$.*

Intuitively, the set of group-traces of a concept class $\mathcal{F}$ represents all possible discriminatory behaviors within $\mathcal{F}$ with respect to the protected groups.

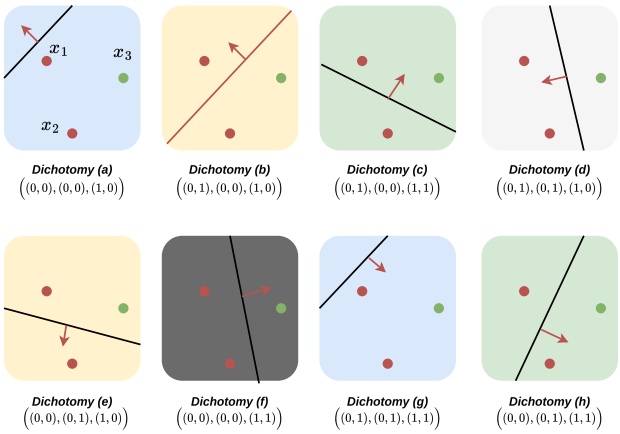


Figure 2: Illustration of SP dimension for the case of non-homogeneous classifiers in $\mathbb{R}^2$

Lemma 3. *For any finite and non-empty sample sets S_0 and S_1 drawn from the first and second protected groups, respectively, the set $\Delta_{\mathcal{F}}^{SP}(S_0, S_1)$ is well-defined.*

Proof To establish the result in Lemma 3, it is sufficient to show that, for any concept class $\mathcal{F}$ of $\mathrm{VC}(\mathcal{F}) = d$, and for any shattered set $S = \{x_1, \cdots, x_d\}$ shattered by $\mathcal{F}$, not all elements of S can belong to the same protected group. By the definition of VC dimension, the total number of dichotomies u for the sample S is 2^d. We assume by contradiction that all elements of S belong to a single protected group. Without loss of generality, we assume that $S \subseteq \mathcal{X}_0$. Let x' be any point from $\mathcal{X}_1$. For all i in $[2^d]$, let c_i denote the concept from $\mathcal{F}$ that realizes the dichotomy u_i. Since $\mathcal{X}_0$ and $\mathcal{X}_1$ form the set of components of $\mathcal{X}$ ($\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$), each c_i can be extended to $\tilde{c_i^0}$ and $\tilde{c_i^1}$, such that $x' \in \tilde{c_i^0}$ and $x' \notin \tilde{c_i^1}$.

Hence, $\mathcal{F}$ realizes 2^{d+1} dichotomies over $S \cup \{x'\}$. In other words, $S \cup \{x'\}$ shatters $\mathcal{F}$. This is a contradiction because $\mathrm{VC}(\mathcal{F}) = d$. ■

Definition 5 (SP dimension)**.** *We say that a sample S SP-shatters by $\mathcal{F}$ when $|\Delta_{\mathcal{F}}^{\mathrm{SP}}(S_0, S_1)| = 2^{|S|} + |S| - 2^{|S_0|} - 2^{|S_1|}$. The SP dimension of a class $\mathcal{F}$ of subsets of $\mathcal{X}$ is*

$$\mathrm{SP}(\mathcal{F}) \triangleq \max_{|S|: S = S_0 \cup S_1} \log_2 \left|\Delta_{\mathcal{F}}^{\mathrm{SP}}(S_0, S_1)\right|.$$

In contrast to Yan and Zhang (2022), whose approach relies on pre-audit model reconstruction via a teaching set and is characterized by classical learning-theoretic complexities (e.g., the disagreement coefficient and VC dimension), the SP dimension of a concept class $\mathcal{F}$ captures only those group-wise dichotomies that exhibit distinct discriminatory behaviors. It achieves this by quotienting out redundant symmetries between protected groups that induce equivalent discrimination patterns within $\mathcal{F}$.

Example 2. *As illustrated in Figure 2, for non-homogeneous linear classifiers in $\mathbb{R}^2$, the VC dimension is 3, which implies $2^3 = 8$ possible dichotomies. Given Lemma 3 requiring at least one point from each protected group to be in the shattered set, the figure depicts this configuration. The same-colored squares in Figure 2 illustrate the same behavior of dichotomies; Consider the green square as an example: it demonstrates a specific behavior with respect to protected groups by assigning a positive label to only one point from the first protected group. This constraint reduces the total number of valid dichotomies from 8 to 5. This is equivalent to computing SP dimension using definition 5: $2^{\mathrm{SP}(\mathcal{F}_2)} = 2^3 + 3 - 2^2 - 2^1 = 5$.*

Theorem 4 (Quantitative characterization)**.** *For any concept class $\mathcal{F}$, the minimum number of samples $m(\mathcal{F}, \epsilon, \delta) > 0$ required to weakly audit SP is lower bounded by $\Omega\left(\frac{\mathrm{SP}(\mathcal{F})}{\epsilon^2}\right)$, while Algorithm 1 requires $\mathcal{O}\left(\frac{32}{\alpha(1-\alpha)\epsilon^2} \max\{\log\frac{2}{\delta}, 2\mathrm{SP}(\mathcal{F}) \log\frac{32e}{\epsilon^2}\}\right)$ samples. Here, $\alpha \in (0,1)$ is the ratio of samples from two protected groups.*

Theorem 4 characterizes the sample complexity bounds for weak auditability when the auditor is given m samples, distributed such that αm samples come from the first group and $(1-\alpha)m$ samples from the second group. For this setting, the theorem establishes both necessary and sufficient conditions: a sample size of $\Omega(\frac{\mathrm{SP}(\mathcal{F})}{\epsilon^2})$ is necessary for weak auditability, while a sample size of $\mathcal{O}\left(\frac{32}{\epsilon^2} \max\{\log\frac{2}{\delta}, 2\mathrm{SP}(\mathcal{F}) \log\frac{32e}{\epsilon^2}\}\right)$ is sufficient.

Corollary 5 (Qualitative characterization)**.** *A concept class $\mathcal{F}$ is agnostic and realizably weak auditable if and only if $\mathrm{SP}(\mathcal{F})$ is finite.*

The proofs are given in Appendix D.3. These bounds establish that weak auditability has a complexity measure similar to learnability. Specifically, the learning problem of a hypothesis class $\mathcal{F}$ has a sample complexity upper bound of $O\left(\frac{\mathrm{VC}(\mathcal{F}) + \log(1/\delta)}{\epsilon^2}\right)$ and a lower bound of $\Omega\left(\frac{\mathrm{VC}(\mathcal{F})}{\epsilon^2}\right)$. These bounds show a tight correspondence with the bounds for weakly auditing. Although, as discussed previously, the SP dimension is upper bounded by the VC dimension. Meaning that the SP dimension allows for fewer possible group behavioural dichotomies compared to classification dichotomies. The sample complexity, however, remains the same up to constant factors.

Proposition 6 (Learnability vs. Auditability)**.** *For*

any $\mathcal{F}$ with finite VC and SP-dimensions, $\mathrm{VC}(\mathcal{F}) \geq \mathrm{SP}(\mathcal{F})$.

Proposition 6 shows that auditing statistical parity has a lower information complexity than learnability. This difference arises because the auditing problem takes advantage of the input space's structure, specifically the symmetry between protected groups. This symmetry makes it easier to analyze how models behave with respect to protected groups, compared to the more complex task of learning the model's behavior across the entire input space. The proof is given in Appendix D.2.

4.2 Strongly auditable classes

We begin by analysing the bounds for strongly auditing the finite hypothesis classes.

Theorem 7 (Strongly auditable finite classes)**.** *Any finite hypothesis class is strongly auditable with a sample complexity* $\mathcal{O}\Big(\max\Big\{\underbrace{\frac{1}{\epsilon^2}\log\frac{|\mathcal{F}|}{\delta}}_{\text{without prospect}}, \underbrace{\frac{1}{\log\frac{1}{\epsilon^2}}\log\frac{|\mathcal{F}|}{\delta}}_{\text{with prospect}}\Big\}\Big)$.

The provided bounds highlight the inherent difficulty of covering the entire prospect class in terms of error when the hypothesis class is finite. This shows the fundamental trade-off of achieving strong auditability (correctness: $\frac{1}{\epsilon^2}$ and completeness: $\frac{2}{\log\frac{2}{\epsilon^2}}$) that imposes an additional cost on sample complexity. The proof is in Appendix E.1. The completeness constraint on strong auditability for infinite model classes presents a fundamental challenge (details in Appendix F). To address it, we introduce a new characterization based on the prospect ratio. We begin by formally defining the prospect class, which serves as building blocks for the prospect ratio definition that follows.

Definition 6 (Prospect class)**.** *Given a distributional property μ, strategic class $\mathcal{F}$, and a parameter $\epsilon \in (0,1)$, the true prospect class with respect to μ is the subclass of models in $\mathcal{F}$ that have the same property μ up to ϵ:* $\mathcal{P}(\mathcal{F},\epsilon) \triangleq \{f \in \mathcal{F} : |\mu(f) - \mu(f^*)| \leq \epsilon\}$.
The empirical prospect class with respect to μ is the subclass of models in $\mathcal{F}$ that have the same property μ on S, up to ϵ: $\hat{\mathcal{P}}(\mathcal{F},\epsilon) \triangleq \{f \in \mathcal{F} : |\hat{\mu}(f) - \mu(f^*)| \leq \epsilon\}$, *where* $f^* \in \arg\min_{f\in\mathcal{F}} |\mu(f) - \mu^*|$, *$\mu^*$ is true property of the black-box model, and ϵ is a threshold.*

Definition 7 (Prospect ratio)**.** *The prospect ratio is the volume of the prospect class compared to the hypothesis class $\mathcal{F}$, i.e.* $r(\epsilon) \triangleq \frac{\mathcal{V}(\hat{\mathcal{P}}(\mathcal{F},\epsilon))}{\mathcal{V}(\mathcal{F})}$. $\mathcal{V} : \mathcal{F} \to \mathcal{R}^+$ *is volume function.*

To address the challenge of infinite prospect classes, we need a systematic way to measure their volume. This can be achieved by reducing the search space to a finite set of representative models through a probability measure over the model class $\mathcal{F}$. Let ν denote this probability measure on $\mathcal{F}$. While the choice of ν affects the prospect ratio's value and is largely arbitrary, it enables us to work with a manageable subset of models. When $\mathcal{F}$ is finite, a natural choice of ν is the uniform distribution, which reduces to the case of strongly auditing finite classes. Let n denote the sample size of models drawn from the strategic class. The prospect ratio is influenced by two distinct sources of uncertainty: (1) uncertainty due to sampling models from the prospect class and (2) uncertainty due to sampling points from protected groups. This induces two types of errors. Let $f_1, \cdots, f_n$ be n models sampled independently from ν, and $\tilde{r}_n(\epsilon) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f_i\in\mathcal{P}(\mathcal{F},\epsilon)}, \hat{r}_n(\epsilon) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f_i\in\hat{\mathcal{P}}(\mathcal{F},\epsilon)}$. We define *the estimator for the prospect ratio* as:

$$\hat{r}_{n,m_0,m_1}(\epsilon) := \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\left|\frac{1}{m_0}\sum_{i=1}^{m_0}\mathbb{1}_{f_k(x_i)=1} - \frac{1}{m_1}\sum_{j=1}^{m_1}\mathbb{1}_{f_k(x'_j)=1}\right| \leq \epsilon}$$

where x_i's and x'_j's are samples from the first and second protected group, respectively.

Theorem 8 (Concentration of prospect ratio)**.** *For all $\epsilon, \upsilon, \tau \in (0,1)$,* $\mathbb{P}\{r(\epsilon-\upsilon)-\tau \leq \hat{r}_{n,m_0,m_1}(\epsilon) \leq r(\epsilon+\upsilon)+\tau\} \geq \left(1-\exp\left\{\frac{-2\upsilon^2 m_0 m_1}{n(m_0+m_1)}\right\}\right)^n (1-\exp(-n\tau^2))^2$.

The prospect ratio estimation error exhibits exponential dependence on the size of the prospect class. While obtaining sufficient samples from each protected group is important for accurate statistical parity estimation, the sample size requirement with respect to models from the hypothesis class is exponentially more critical for achieving strong auditability (Appendix E.2).

Auditable but Unlearnable: Infinite VC Classes. The following proposition 9 provides a negative answer to the second question posed in the introduction. This question is equivalent to asking whether we can reduce an infinite set of dichotomies to a finite set that captures all distinct behaviors with respect to protected groups.

Proposition 9. *Any class of infinite VC dimension can not be weakly or strongly auditable.*

The proof is given in Appendix E.3.

5 NUMERICAL EXPERIMENTS

As discussed in Section 4.2, auditing an infinite hypothesis class is generally intractable. To address this, we adopt the sampling-based approximation described therein: we draw a fixed number of hypotheses to construct a finite representative subset. This subset preserves the geometric structure of the underlying strategic class while enabling the empirical evaluation of SP.

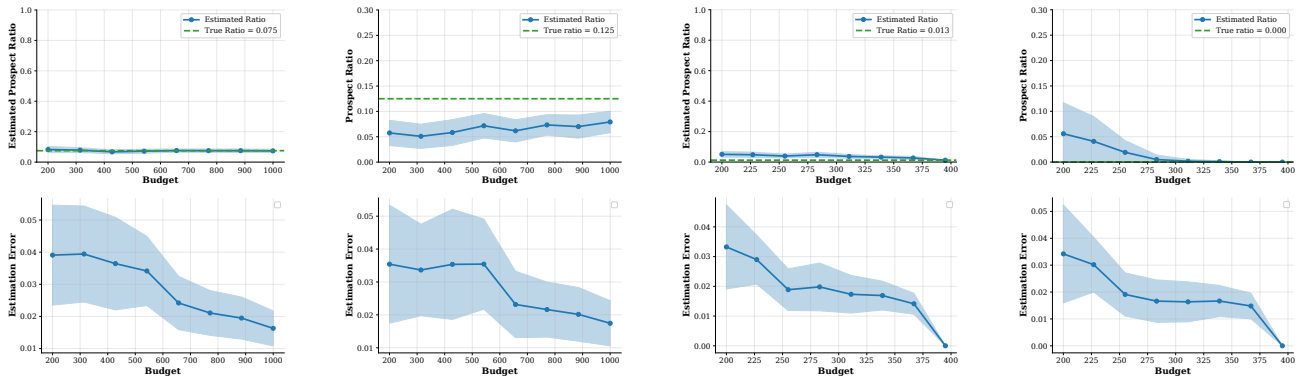


Figure 3: Comparison of errors in statistical parity estimation and prospect ratio across different sample sizes.

5.1 Experimental Setup

Within this finite approximation, we evaluate ERMP (Algorithm 1) for strong auditability according to conditions (i) and (ii) in Definition 3. While Yan and Zhang (2022) evaluates the prospect class using its *diameter*, $\max_{f,f' \in \mathcal{P}(\mathcal{F},\epsilon)} \mu(f) - \mu(f')$ which captures the worst-case spread of statistical parity values consistent with the observed data, a small diameter may arise simply because the prospect class contains only a single model. Such a scenario would trivially satisfy fairness but violate the *completeness* requirement of strong auditing, namely that the auditor must consider a sufficiently rich set of compliant models. To address this limitation, we complement the diameter-based analysis with an evaluation of condition (ii) using the *estimated prospect ratio*, which better reflects diversity of the prospect class beyond statistical error.

We evaluate our approach on two standard fairness benchmarks: the COMPAS dataset (Angwin et al. 2016), where groups are defined as Caucasian and non-Caucasian, and the Student Performance dataset, with groups defined as Female and Male. The true black-box model is a logistic regression classifier with ℓ_2 regularization, trained on the original labels using scikit-learn's default solver (Pedregosa et al. 2011). We consider two strategic model updates: one replacing the original model with a multi-layer perceptron (MLP), and another with a random forest.

5.2 Audit Fidelity Evaluation

We assess the fidelity of our auditing procedure along two dimensions: (a) *Prospect class comprehensiveness*: We measure the prospect ratio, the fraction of hypotheses that match the true model's statistical parity, and observe its convergence across increasing labeling budgets. As shown in the first row of Figure 3, the estimated ratio converges to the ground-truth ratio (computed on the full dataset), empirically validating that our algorithm reliably captures the true prospect class in the strong auditing setting. (b) *Prospect class correctness*: We evaluate whether models in the prospect class indeed exhibit statistical parity equivalent to the true model. The second row of Figure 3 shows that the statistical parity estimation error for prospect models decreases with budget and remains consistently below our tolerance threshold $\epsilon = 0.005$, confirming the correctness of our prospect class construction.

5.3 Audit's Runtime Evaluation

EPRM does not scales with the labeling budget, particularly for simpler strategic updates (i.e., random forests) requiring only **3 ms per sample** on average (see Table 1 and Figure 5). This runtime accounts for both statistical parity estimation and prospect class construction. Additional details and hardware specifications are provided in Appendix G.

6 DISCUSSION & FUTURE WORK

We characterize the class of allowable strategic model updates by model owners — namely, those that preserve the value of the audited property. We establish a necessary and sufficient condition in terms of the SP dimension, a complexity measure strictly weaker than VC dimension. Our results suggest three natural directions for future work: (i) extending the framework to interactive settings via sequential, property-aware complexity measures; (ii) embedding audited properties directly into model architectures to obtain optimal, property-preserving predictors; and (iii) developing manipulation-proof audit definitions that are concept-class-agnostic and dimension-free.

Acknowledgments

This work is supported by the Regalia project of Inria and French Ministry. We also acknowledge the ANR

JCJC project REPUBLIC (ANR-22-CE23-0003-01), the PEPR project FOUNDRY (ANR23-PEIA-0003), and the Inria-ISI Kolkata associate team SeRAI for partially supporting the project.

References

Mateo Aboy, Timo Minssen, and Effy Vayena. Navigating the eu ai act: implications for regulated digital medical products. *npj Digital Medicine*, 7(1): 237, 2024.

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

Ayoub Ajarra, Bishwamittra Ghosh, and Debabrota Basu. Active fourier auditor for estimating distributional properties of ml models. *arXiv preprint arXiv:2410.08111*, 2024.

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

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

Shai Ben-David, Alon Itai, and Eyal Kushilevitz. Learning by distances. In *COLT*, pages 232–245, 1990.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024.

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

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

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

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

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Yi Hao and Ping Li. Bessel smoothing and multi-distribution property estimation. In *Proceedings of Thirty Third Conference on Learning Theory (COLT)*, 2020.

Yi Hao and Alon Orlitsky. Data amplification: Instance-optimal property estimation. In *International Conference on Machine Learning (ICML)*, 2020.

Yi Hao, Alon Orlitsky, Ananda T. Suresh, and Yihong Wu. Data amplification: A unified and competitive approach to property estimation. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Daniel Hsu, Jizhou Huang, and Brendan Juba. Distribution-specific auditing for subgroup fairness. In *5th Symposium on Foundations of Responsible Computing (FORC 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024.

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

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

Erwan Le Merrer, Ronan Pons, and Gilles Trédan. Algorithmic audits of algorithms, and the law. *AI and Ethics*, pages 1–11, 2023.

Jinxin Liu, Michele Nogueira, Johan Fernandes, and Burak Kantarci. Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems. *IEEE Communications Surveys & Tutorials*, 24(1):123–159, 2021.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363, 2018.

Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

Christian Montag and Michèle Finck. Successful implementation of the eu ai act requires interdisciplinary efforts. *Nature Machine Intelligence*, 6(12): 1415–1417, 2024.

Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. *Advances in Neural Information Processing Systems*, 35:37458–37470, 2022.

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

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

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

Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Auditing black-box prediction models for data minimization compliance. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

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

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

Yining She, Sumon Biswas, Christian Kästner, and Eunsuk Kang. Fairsense: Long-term fairness analysis of ml-enabled systems. *arXiv preprint arXiv:2501.01665*, 2025.

Alice Silva. Using data mining to predict secondary school student performance. 2008.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. In *The Electronic Journal of Statistics (EJS)*, 2012.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Briana Vecchione, Karen Levy, and Solon Barocas. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

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

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

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

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]: Code, with instructions on how to navigate the repository and how to run the experiments, is provided at : `https://anonymous.4open.science/r/Auditors-with-prospects-050F/`
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators.[Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Extended Related Works

A.1 Estimating Distributional Properties

The auditing of properties of ML models has been explored in various contexts. For example, Rastegarpanah et al. (2021) investigated model instability under feature imputation in a black-box setting, with guarantees on data minimization. More relevant to our work, several studies have focused on auditing group fairness to examine discriminatory behavior with respect to protected groups. In particular, Kearns et al. (2018) studied the statistical parity audits through a reduction to weak agnostic learning, where auditing is defined as verifying whether SP exceeds a given threshold. Building on this, Hsu et al. (2024) recently explored a specific case of this reduction, focusing on auditing SP under Gaussian feature distributions for homogeneous halfspace subgroups, and demonstrated the problem's computational hardness. Similarly, Chugg et al. (2023) examined the same verification problem within the statistical framework of hypothesis testing. More closely related to our work, Yan and Zhang (2022) examined auditing statistical parity via estimation instead of verification. In their approach, manipulation-proof constraints are enforced through active sampling. However, restricted to a finite hypothesis class, their method relies on reconstructing the model before plugging in the estimator. This limitation leads to the potential omission of hypotheses that could expand the size of the manipulation subclass, reducing the overall effectiveness of the auditing process.

In contrast to previous approaches, a variant of this problem has been investigated to audit distributional properties, where the focus is on estimating characteristics of an unknown distribution. Examples of such properties include Shannon entropy, the size of the distribution's support, support coverage, and various distribution metric distances (such as KL divergence and other divergence measures). Common approaches are often based on plug-in estimators that approximate the unknown distribution, typically requiring a logarithmic-factor increase in additional samples. Recent works by Hao and Orlitsky (2020); Hao et al. (2018) have explored this challenge within the realm of discrete distributions, incorporating smoothness assumptions and proposing an estimator that amplifies data relative to the empirical estimator by a factor $\sqrt{\log n}$. Hao and Li (2020) extended the problem to the multi-distribution setting to estimate discrete distribution properties over a class of discrete distributions over $[k]$ considering their mixtures (mixing strategies) and maintaining smoothness conditions with sample complexity $\mathcal{O}(\frac{k}{\epsilon^3\sqrt{\log k}})$. Sriperumbudur et al. (2012) explores integral probability metrics for continuous distributions, expanding upon various known distance metrics between distributions, such as total variance, Wasserstein distance, among others. Sriperumbudur et al. (2012) further extended their analysis to kernelized distances, enriching the understanding of distance measures in the context of continuous distributions.

A.2 Auditing with Prospects: Connections to Rashomon Sets

Recent work has studied the complexity of learning "simple" models, which carry nuanced meanings across ML communities. For example, in healthcare applications, simplicity often refers to developing explainable models that address the black-box inexplainability problem Rudin (2019). Alternatively, in fairness-aware machine learning, simplicity can mean identifying fair models within an equally accurate model class Agarwal et al. (2018). This phenomenon, known as the Rashomon effect, aims to explore multiple perspectives of the joint dataset distribution, revealing different truths. The Rashomon set represents a collection of models that achieve comparable performance but differ in their explanations or underlying patterns. The Rashomon ratio, introduced by (Semenova et al. 2022), quantifies this effect by measuring the volume of the Rashomon class relative to the hypothesis class.

Definition 8 (Rashomon set)**.** *Given $\epsilon > 0$, a dataset S, a hypothesis class $\mathcal{H}$ and a loss function ℓ,*

1. *The true Rashomon set is:*

$$\mathcal{R}_{\mathcal{D}}(\mathcal{H}, \epsilon) \triangleq \{h \in \mathcal{H} : \mathcal{L}_{\mathcal{D}}(h) \leq \arg\min_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h') + \epsilon\}$$

2. *The empirical Rashomon set is:*

$$\hat{\mathcal{R}}_{S}(\mathcal{H}, \epsilon) \triangleq \{h \in \mathcal{H} : \hat{\mathcal{L}}_{S}(h) \leq \arg\min_{h' \in \mathcal{H}} \hat{\mathcal{L}}_{S}(h') + \epsilon\}$$

where $\mathcal{L}_{\mathcal{D}}$ *and* $\hat{L}_S$ *are the true and empirical risk respectively, defined as* $\mathcal{L}_{\mathcal{D}} = \mathop{\mathbb{E}}_{(x,y)\mathcal{D}}\{l(h,(x,y))\}$, $\hat{L}_S = \frac{1}{|S|}\sum_{(x,y)\in S} \ell(h,(x,y))$.

While Rashomon sets and prospect classes both explore sets of models that maintain specified performance properties, such as bounded learning error or auditing error, their theoretical foundations differ significantly. Rashomon sets have primarily been studied in the context of finite hypothesis classes Semenova et al. (2022), where the central theoretical challenge is not the information-theoretic complexity of learning [2], but rather quantifying the relative simplicity of learning with respect to the empirical data distribution. This simplicity is formally captured by the Rashomon ratio, which measures the proportion of models achieving a specified performance threshold. Consequently, practitioners typically focus on selecting a single 'simple' model from the Rashomon set, rather than characterizing its complete structure or geometric properties. In contrast, prospect classes, central to manipulation-proof auditing theory, require explicit characterization of all models that satisfy the auditing criteria. This comprehensive enumeration requirement arises from the need to reason about all possible ways a model owner might manipulate the pre-audit model while maintaining acceptable performance.

The distinction highlights that prospect classes and Rashomon sets, despite their structural similarities in characterizing sets of property-preserving models, serve fundamentally different theoretical objectives: while Rashomon sets focus on quantifying model simplicity within finite hypothesis classes via the Rashomon ratio, prospect classes are concerned with the information-theoretic complexity of learning the entire set of property-preserving models in potentially infinite hypothesis spaces.

B Additional Auditing Scenarios

In statistical learning, we study models that aim to replicate the behavior of the data-generating distribution. While the ultimate goal is to approximate this behavior, stronger theoretical guarantees, such as those under the *realizability assumption*, are often obtained by assuming that some model within the hypothesis class exactly captures the joint distribution of the data. This perspective can be reframed in an auditing context: instead of seeking a model with minimal error, we audit models whose error is close to that of a given reference model (the model under audit)[3]. In other words, the task becomes estimating the *true error* of the reference model—not necessarily the minimum error achievable over the hypothesis class, but rather the smallest error attainable by models whose performance is comparable to that of the fixed reference model. In this setting, the regulator, denoted by $\mathfrak{R}$ (the auditor), operates in a black-box regime: the internal structure of the learning algorithm (illustrated by the left part in Figure 1) is inaccessible. The auditor $\mathfrak{R}$ is limited to a single pass over data sampled from the joint distribution $\mathcal{D}$ and must output a *prospect model* (in the weakly auditable framework) or a *prospect class* (in the strongly auditable framework) that approximates the true error of the black-box model under audit.

Although this black-box auditing scenario may not correspond directly to real-world applications, it serves as a theoretical generalization of agnostic learning to other model properties that are practically relevant for regulatory compliance. This abstraction extends Algorithm 1 to diverse auditing contexts and underscores the robustness of agnostic frameworks, demonstrating their relevance in regulatory settings where access to a model's internal parameters or architecture is restricted.

[2] as this is already characterized by the VC dimension.

[3] Note that this error need not be zero.

An analogy for PAC-audit: Bridging concepts. As explained above, in the context of statistical learning, the property μ corresponds to the true error. Here, the ground truth is zero (i.e., $\mu(\mathcal{D}) = 0$) since the data-generating distribution $\mathcal{D}$ itself encodes the correct labeling. In this setting, the prediction error and the learning error are tightly coupled: for every hypothesis $f \in \mathcal{F}$, the (true) auditing error $|\mu(f) - \mu(\mathcal{D})|$ coincides with the standard true learning error $\mathcal{L}_{\mathcal{D}}(f) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$. Thus, this is a special case in which the auditing discrepancy can be expressed directly as a joint functional of the hypothesis and the distribution, i.e., $|\mu(f) - \mu(\mathcal{D})| = \mu(f, \mathcal{D})$.

In summary, classical PAC learning outputs a single hypothesis that, with high probability, has true error close to the minimum, along with a provable generalization bound. This can be viewed as a special case of weakly audit (Definition 2), which generalizes the framework by replacing the error functional with a distributional property of interest μ.

B.1 Loss Functions in The Context of Audits

Rather than estimating the audited property directly, our framework aims to emulate a black-box reference model with respect to a fixed target property. Conventional estimation approaches focus solely on estimation error, overlooking a key constraint introduced earlier: the model owner must adapt their model to satisfy strategic objectives, which we term prospects. To address this, we propose a new class of loss functions explicitly tied to the property of interest. Below, we present a general definition of these loss functions, formulated in terms of multiple populations encoded within the property.

Definition 9. *For any auditing problem, let $\{\mathcal{X}_i\}_{i\in I}$ denote a feature partition of $\mathcal{X}$, μ a distributional property defined over this partition, f^* the model under audit, and f the model produced by the auditor.*

- *The audit loss is defined as $\ell_\mu : \mathcal{Y}^{2|I|} \longrightarrow [0, 1]$.*
- *The audit risk is defined as:*

$$\mathcal{E}_\mu(f) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell_\mu((f^*(x_i), y_i)_{i\in I}, (f(x_i), y_i)_{i\in I})|x_i \in \mathcal{X}_i]$$

- *Let $S = \bigcup_{i\in I} S_i$ denote a set of instances sampled i.i.d from $\mathcal{D}$, where $S_{i_{|\mathcal{X}}} \subseteq \mathcal{X}_i$. For all $i \in I$, let m_i denote the cardinal of S_i and $S_i = \{(x_i^j, y_i^j)\}_{j=1}^{m_i}$. Let $m = \prod_{i\in I} m_i$. The empirical auditing risk with respect to S is*

$$\hat{\mathcal{E}}_\mu(f) \triangleq \frac{1}{m}\sum_{j=1}^{m} \ell_\mu((f(x_i^j), y_i^j)_{i\in I})$$

In the audit loss function defined above, the comparison is made pointwise between the output of the model under audit and that of the auditor's reference model (i.e., the prospect model), evaluated over the relevant population(s). Notably, the loss function inherently depends on the distributional property of interest, as it encodes distinctions across multiple populations. This generalized formulation extends the classical PAC learning framework—which is typically restricted to a single data distribution—to settings involving multiple populations. Consequently, it enables learning and evaluation with respect to properties that are defined jointly over more than one population. As we now illustrate through concrete examples, the audit problem naturally reduces to a risk minimization problem analogous to those central to statistical learning theory.

B.2 Weakly Audits for General Distributional Properties

In this section, we demonstrate that Definition 9 applies to a broad class of distributional properties (examples in section 2) in the weak auditing setting (definition 2). We further show that minimizing the expected audit loss (i.e., the audit risk) yields models whose predictions are statistically close to the true property values, thereby extending classical estimation problems to the learning-theoretic setting of approximating audited properties.

In the following, f^* denotes the model under audit, which is accessible only via a finite set of labeled samples, and f denotes the auditor's prospect model.

B.2.1 Learning Error

As discussed previously, standard learning problems are evaluated in a single population (the dataset). Consequently, in Definition 9, we have $|I| = 1$. The goal is to approximate the model under audit f^* with the auditor's prospect model f in the same data distribution $\mathcal{D}$. In this setting, the audit loss (for binary classifiers [4]) is defined as:

$$\ell((f(x), y), (f^*(x), y)) = |\mathbb{1}_{f(x)\neq y} - \mathbb{1}_{f^*(x)\neq y}|$$

The corresponding audit risk is the following.

$$\mathcal{E}(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell((f(x), y), (f^*(x), y))]$$

Proposition 10. *Audit Risk minimization implies weakly audit of the learning error.*

Proof Let $(\epsilon, \delta) \in (0, 1)^2$, S denotes an audit set of size m, and $f_S = f$ a prospect model, by the Jensen's inequality:

$$\begin{aligned}
|\mu(f_S) - \mu^*| &= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{f_S(x)\neq y} - \mathbb{1}_{f^*(x)\neq y}] \right| \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[|\mathbb{1}_{f_S(x)\neq y} - \mathbb{1}_{f^*(x)\neq y}|] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell((f_S(x), y), (f^*(x), y))] \\
&= \mathcal{E}(f_S)
\end{aligned}$$

This is true for every audit set S sampled i.i.d from $\mathcal{D}$. Therefore, we have:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[|\mu(f_S) - \mu^*| > \epsilon\Big] \leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[\mathcal{E}_\mu(f_S) > \epsilon\Big]$$

$$\textit{C/C:} \qquad \forall(\epsilon, \delta) \in (0, 1)^2, \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[\mathcal{E}_\mu(f_S) > \epsilon\Big] \leq \delta \implies \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[|\mu(f_S) - \mu^*| > \epsilon\Big] \leq \delta$$

■

One may ask whether the VC dimension still characterizes the information complexity of weak auditability. The answer is affirmative: the classic proof, based on a union bound over all hypotheses in a class of finite VC dimension, carries over unchanged, since the audit loss remains a binary function of model predictions.

B.2.2 Statistical Parity

Statistical parity is a group fairness property defined with respect to a binary protected attribute, comparing outcomes across the two resulting populations (i.e., $|I| = 2$). Accordingly, the audit loss function operates on pairs of predictions, one from each group, and is defined over $\mathcal{Y}^2$. Formally, for inputs x_0 and x_1 drawn from each protected group, the audit loss is:

$$\ell\Big((f(x_0), y_0), (f^*(x_1), y_1)\Big) \triangleq \Big|\big(\mathbb{1}_{f(x_0)=y_0} - \mathbb{1}_{f(x_1)=y_1}\big) - \big(\mathbb{1}_{f^*(x_0)=y_0} - \mathbb{1}_{f^*(x_1)=y_1}\big)\Big|$$

[4] The same results generalize to different models, including regressions via the convexity property of the squared loss.

Where y_0 and y_1 are often assumed constants equal to one, and interpret statistical parity as equality of positive prediction rates across groups. The loss is zero for a given pair if f and f^* exhibit the same group-wise disparity: that is, the difference in the prediction between groups of the model under audit matches that of the prospect model. The corresponding audit risk is the following.

$$\mathcal{E}(f) = \mathop{\mathbb{E}}_{((x_0,y_0),(x_1,y_1))\sim\mathcal{D}^2}\Big[\ell\Big((f(x_0),y_0),(f^*(x_1),y_1)\Big)\Big|(x_0,x_1)\in\mathcal{X}_0\times\mathcal{X}_1\Big]$$

Proposition 11. *Audit Risk minimization implies weakly audit of the statistical parity.*

The proof follows similarly to that of Proposition 10.

Multi-group Fairness. The preceding arguments extend naturally to the multi-group fairness property, which involves p protected groups. This property is defined as the maximum discrepancy in positive prediction rates across all pairs of groups:

$$\mu(f) = \max_{(i,j)\in I} |\mathop{\mathbb{P}}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)=1|x\in\mathcal{X}_i] - \mathop{\mathbb{P}}_{x\sim\mathcal{D}_{|\mathcal{X}}}[f(x)=1|x\in\mathcal{X}_j]|.$$

where $|I| = p$. The audit loss and audit risk are the ones given in definition 9.

B.2.3 Learning Stability

For a model trained on a distribution $\mathcal{D}_{\text{src}}$ and deployed on $\mathcal{D}_{shift}$. We defined the learning stability of f as $\mu(f,\mathcal{D}_{\text{src}},\mathcal{D}_{\text{shift}}) \triangleq \Big|\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}_{\text{src}}}[f(x)\neq y] - \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}_{\text{shift}}}[f(x)\neq y]\Big|$. Here, the property is defined with respect to two distinct data distributions: a source distribution and a shifted distribution. We demonstrate that the earlier theoretical guarantees generalize to such settings, where audited properties depend on multiple distributions rather than on subpopulations of a single distribution. The audit loss function operates on pairs of predictions from two models (the prospect model and the model under audit) evaluated on samples from two distinct distributions. Specifically, for a given input, the loss takes as arguments the outputs of both models on both distributions, resulting in a function defined over $\mathcal{Y}^4$. Formally, the loss is defined as:

$$\ell\Big(((f(x),y),(f(\tilde{x}),\tilde{y}));((f^*(x),y),(f^*(\tilde{x}),\tilde{y}))\Big) \triangleq \Big|\big(\mathbb{1}_{f(x)\neq y} - \mathbb{1}_{f(\tilde{x})\neq\tilde{y}}\big) - \big(\mathbb{1}_{f^*(x)\neq y} - \mathbb{1}_{f^*(\tilde{x})\neq\tilde{y}}\big)\Big|$$

And the audit risk for learning stability is:

$$\mathcal{E}(f) = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\mathcal{D}_{\text{src}}\\(\tilde{x},\tilde{y})\sim\mathcal{D}_{\text{shift}}}}\Big[\ell\Big(\big((h(x),y),(h(\tilde{x}),\tilde{y})\big),\big((h^*(x),y),(h^*(\tilde{x}),\tilde{y})\big)\Big)\Big]$$

As in the earlier setting, our results extend to properties defined over distinct distributions:

Proposition 12. *Audit risk minimization implies weakly auditing of the stability of learning.*

Proof Let $(\epsilon,\delta)\in(0,1)^2$. We follow the same previous steps:

$$\begin{aligned}
\mathcal{E}(f) &= \mathop{\mathbb{E}}_{\substack{(x,y)\sim\mathcal{D}_{\text{src}}\\(\tilde{x},\tilde{y})\sim\mathcal{D}_{\text{shift}}}}\Big[\ell\Big(\big((f(x),y),(f(\tilde{x}),\tilde{y})\big),\big((f^*(x),y),(f^*(\tilde{x}),\tilde{y})\big)\Big)\Big]\\
&= \mathop{\mathbb{E}}_{\substack{(x,y)\sim\mathcal{D}_{\text{src}}\\(\tilde{x},\tilde{y})\sim\mathcal{D}_{\text{shift}}}}\Big[\Big|\big(\mathbb{1}_{f(x)\neq y} - \mathbb{1}_{f(\tilde{x})\neq\tilde{y}}\big) - \big(\mathbb{1}_{f^*(x)\neq y} - \mathbb{1}_{f^*(\tilde{x})\neq\tilde{y}}\big)\Big|\Big]\\
&\geq \Big|\mathop{\mathbb{E}}_{\substack{(x,y)\sim\mathcal{D}_{\text{src}}\\(\tilde{x},\tilde{y})\sim\mathcal{D}_{\text{shift}}}}\Big[\big(\mathbb{1}_{f(x)\neq y} - \mathbb{1}_{f(\tilde{x})\neq\tilde{y}}\big) - \big(\mathbb{1}_{f^*(x)\neq y} - \mathbb{1}_{f^*(\tilde{x})\neq\tilde{y}}\big)\Big]\Big|\\
&= |\mu(f) - \mu^*|
\end{aligned}$$

Hence,

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[|\mu(f_S)-\mu^*|>\epsilon\Big] \leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[\mathcal{E}_\mu(f_S)>\epsilon\Big]$$

$$\textit{C/C:} \qquad \forall(\epsilon,\delta)\in(0,1)^2 : \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[\mathcal{E}_\mu(f_S)>\epsilon\Big]\leq\delta \implies \mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[|\mu(f_S)-\mu^*|>\epsilon\Big]\leq\delta$$

■

B.2.4 Robust risk

For any x in $\mathcal{X}$, let $\mathcal{U}(x)$ denote the set of admissible perturbations of x. The robust risk of a classifier f defined as $\mu(f,\mathcal{D},\mathcal{U}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[\sup_{z\in\mathcal{U}(x)} \mathbb{1}_{f(z)\neq y}\Big]$. Similarly to previous discussed properties, we extend the audit framework to the property of robust risk with respect to arbitrary perturbation sets. Extending the audit framework to this notion of robust risk, defined with respect to arbitrary perturbation sets, we introduce the following auditing loss:

$$\ell_{\mathcal{U}}((h(\mathcal{U}(x)),y),(h^*(\mathcal{U}(x)),y)) = \sup_{z\in\mathcal{U}(x)} |\mathbb{1}_{h(z)\neq y} - \mathbb{1}_{y^*\neq y}|$$

And the corresponding audit risk:

$$\mathcal{E}_{\mathcal{U}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}((h(\mathcal{U}(x)),y),(h^*(\mathcal{U}(x)),y))]$$

Proposition 13. *For any set of perturbations $\mathcal{U}$, risk minimization implies weakly auditing robust risk.*

An intriguing question arises: which strategic classes (hypothesis classes) $\mathcal{F}$ are weakly auditable with respect to an arbitrary perturbation set $\mathcal{U}$? The answer depends crucially on the interaction between the perturbation set $\mathcal{U}$ and the strategic class $\mathcal{F}$.

We conjecture that the capacity measure $\mathfrak{D}_{\mathcal{U}}(\mathcal{H})$, introduced by Montasser et al. (2022), provides a meaningful characterization of the complexity of adversarially robust learning. This measure is defined as

$$\mathfrak{D}_{\mathcal{U}}(\mathcal{H}) = \max\left\{ n\in\mathbb{N}\cup\{\infty\} \;\middle|\; \begin{array}{l} \exists \text{ a finite subgraph } G=(V,E) \text{ of } G^{\mathcal{U}}_{\mathcal{H}} \text{ such that} \\ \forall \text{ orientations } \mathcal{O} \text{ of } G,\ \exists\, v\in V \text{ with } \mathrm{outdeg}(v;\mathcal{O})\geq \frac{n}{3}. \end{array} \right\}$$

Conjecture 14. *There exists an (ϵ,δ)- weakly auditor for the robust risk auditing problem $\langle\mathcal{X},\mathcal{Y},\mathcal{H},\mathcal{P},\mu_{\mathcal{U}},\ell_{\mathcal{U}}\rangle$ if and only if a complexity measure $\mathfrak{D}_{\mathcal{U}}(\mathcal{H})$ if finite.*

The conjecture says that $\mathfrak{D}_{\mathcal{U}}(\mathcal{F})$ effectively captures the intrinsic difficulty of robust risk weakly audit, thereby extending the characterization from robust statistical learning to the weakly audit of the robust risk.

C Proof of strategic lemma

Before proceeding with the proof, we restate the lemma 1 for clarity.

Lemma 1 (Strategic Lemma)**.** *Let $\epsilon,\delta\in(0,1)$, $m:(0,1)^2\to\mathbb{N}$. Suppose that the following holds:*

- ***Estimation accuracy.*** *$\mathcal{A}$ outputs f_S from $\mathcal{F}$: $\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\Big[|\hat{\mu}_S(f_S)-\widehat{OPT}(S,\mathcal{F})|>\frac{\epsilon}{3}\Big]<\frac{\delta}{2}$.*

- ***Uniform convergence.*** *$\mathcal{F}$ verifies* $\mathbb{P}_{S\sim\mathcal{D}^m}\left[\exists f \in \mathcal{F}, |\mu_{\mathcal{D}}(f) - \hat{\mu}_S(f)| > \frac{\epsilon}{3}\right] < \frac{\delta}{2}$.

Then, $\mathcal{A}$ is (ϵ, δ)-weak auditor for statistical parity.

Our goal is to select, from the strategic class $\mathcal{F}$, a prospect model that minimizes estimation error. This involves two key ingredients: (i) an algorithmic property, the use of an empirical audit risk minimizer to select the model. And (ii) a statistical property, uniform convergence which guarantees that the empirical performance generalizes to unseen data.

Proof

To unify the audit problems, we prove the result for an arbitrary distributional property; the claim for statistical parity then follows immediately. Let $\mathcal{D}$ be a distribution in $\mathcal{X} \times \mathcal{Y}$, S be a set of samples from $\mathcal{D}$, and $f_S \in \mathcal{A}(S)$ denote the prospect model produced by the auditor $\mathcal{A}(S)$. By the triangle inequality,

$$|\mu_{\mathcal{D}}(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| \leq |\mu_{\mathcal{D}}(f_S) - \hat{\mu}_S(f_S)| + |\hat{\mu}_S(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| + |\text{opt}(\mathcal{D}, \mathcal{F}) - \text{opt}(\mathcal{D}, \mathcal{F})|$$

This inequality is verified for any S sampled i.i.d from $\mathcal{D}^n$, and hence,

$$\mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \text{OPT}(\mathcal{D}, \mathcal{F})| \leq \epsilon\right] \geq \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \hat{\mu}_S(f_S)| \leq \frac{\epsilon}{3}\right] + \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\hat{\mu}_S(f_S) - \text{OPT}(\mathcal{D}, \mathcal{F})| \leq \frac{\epsilon}{3}\right] + \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\text{OPT}(\mathcal{D}, \mathcal{F}) - \text{opt}(\mathcal{D}, \mathcal{F})| \leq \frac{\epsilon}{3}\right] - 2$$

Equivalently,

$$\mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| > \epsilon\right] \leq \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \hat{\mu}_S(f_S)| > \frac{\epsilon}{3}\right] + \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\hat{\mu}_S(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| > \frac{\epsilon}{3}\right] + \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\text{opt}(\mathcal{D}, \mathcal{F}) - \text{opt}(\mathcal{D}, \mathcal{F})| > \frac{\epsilon}{3}\right]$$

On the other hand,

$$\begin{cases} |\mu_{\mathcal{D}}(f_S) - \hat{\mu}_S(f_S)| > \frac{\epsilon}{3} & \implies \exists h \in \mathcal{F}, |\mu_{\mathcal{D}}(h) - \hat{\mu}_S(h)| > \frac{\epsilon}{3} \\ |\text{opt}(\mathcal{D}, \mathcal{F}) - \hat{\text{opt}}(S, \mathcal{F})| > \frac{\epsilon}{3} & \implies \exists h \in \mathcal{F}, |\mu_{\mathcal{D}}(h) - \hat{\mu}_S(h)| > \frac{\epsilon}{3} \end{cases}$$

By uniform convergence property of $\mathcal{F}$,

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| > \epsilon\right] &< \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\mu_{\mathcal{D}}(f_S) - \hat{\mu}_S(f_S)| > \frac{\epsilon}{3}\right] + \mathbb{P}_{S\sim\mathcal{D}^n}\left[|\hat{\mu}_S(f_S) - \text{opt}(\mathcal{D}, \mathcal{F})| > \frac{\epsilon}{3}\right] \\ &< \delta\end{aligned}$$

The lemma holds for any distributional property. In particular, it applies to statistical parity. ■

D Proofs for weakly auditable classes

D.1 All finite classes are weakly auditable (Proof of Theorem 2)

We begin by restating Theorem 2:

Theorem 2 (Agnostic weak auditability). *If $\mathcal{F}$ is a finite hypothesis class, than $\mathcal{F}$ is weakly auditable, with respect to statistical parity, for any distribution on $\mathcal{X} \times \mathcal{Y}$ with a sample complexity $\mathcal{O}\Big(\Big\lceil \frac{18}{\epsilon^2} \log \frac{8|\mathcal{F}|}{\delta} \Big\rceil\Big)$.*

Proof For $i \in \{0, 1\}$, let m_i denote the sample size of the i-th protected group.

Since $\mathcal{F}$ is finite, it is sufficient to show the uniform convergence property with respect to statistical parity and deduce the result using Lemma 1.

From Lemma 24, we have

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big[|\mu_{\mathcal{D}}(f) - \hat{\mu}_S(f)| \geq \frac{\epsilon}{3}\Big] \leq \exp\{-\frac{m_0 m_1 \epsilon^2}{18(m_0+m_1)}\}$$

By triangle inequality, we have for all $f \in \mathcal{F}$:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big[|\mu_{\mathcal{D}}(f) - \widehat{OPT}(S,\mathcal{F})| \geq \frac{\epsilon}{3}\Big] \leq \exp\{-\frac{m_0 m_1 \epsilon^2}{18(m_0+m_1)}\} \tag{1}$$

By the union bound over finite strategic class $\mathcal{F}$, and Claim 21, the right part in inequality 1 is upper bounded by δ if the sample complexity m verifies $m = \mathcal{O}\Big(\Big\lceil \frac{18}{\epsilon^2} \log \frac{8|\mathcal{F}|}{\delta} \Big\rceil\Big)$

■

D.2 Relationship between VC and SP dimensions

The SP dimension of a strategic class $\mathcal{F}$ is the largest integer $m = m_0 + m_1$ such that there exists a set of m_0 resp. m_1) points from the first protected group (resp. second protected group) that can be SP-shattered (i.e., for every possible pair labeling (dichotomy) in a certain set of strategically feasible labels, there exists a classifier $f \in \mathcal{F}$ that induces that statistical parity value after the labels are revealed.

By definition, the set of SP-realizable dichotomies on any sample is a subset of the set of all dichotomies realizable by $\mathcal{F}$ in the standard (shattering) definition. Since the VC dimension is the maximum m such that all 2^m dichotomies are realizable, it follows that if $\mathcal{F}$ cannot shatter a set of size m in the VC sense, it cannot SP-shatter it either. Therefore, the SP dimension cannot exceed the VC dimension.

D.3 Weakly auditability does not extend to all infinite classes

First, we prove a technical lemma showing that SP dimension is well defined.

Lemma 15 (Lemma 3). *For any finite subsets S_0 and S_1 drawn from the first and second protected groups respectively, the set $\Delta_{\mathcal{C}}^{SP}(S_0, S_1)$ is well defined.*

Proof To establish the result in Lemma 3, it is sufficient to show that, for any concept class $\mathcal{C}$ of $\mathrm{VC}(\mathcal{F}) = d$, and for any shattered set $S = \{x_1, \cdots, x_d\}$ shattered by $\mathcal{F}$, not all elements of S can belong to the same protected group.

By the definition of VC dimension, the total number of dichotomies for the sample S is 2^d. We assume by contradiction that all elements of S belong to a single protected group. Without loss of generality, we assume that $S \subseteq \mathcal{X}_0$. Let x' be any point from $\mathcal{X}_1$.

For all i in $[2^d]$, let c_i denote the concept from $\mathcal{C}$ that realizes the dichotomy u_i. Since $\mathcal{X}_0$ and $\mathcal{X}_1$ form the set of components of $\mathcal{X}$ ($\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$), each c_i can be extended to $\tilde{c_i^0}$ and $\tilde{c_i^1}$, such that $x' \in \tilde{c_i^0}$ and $x' \notin \tilde{c_i^1}$.

Hence, $\mathcal{C}$ realizes 2^{d+1} dichotomies over $S \cup \{x'\}$. In other words, $S \cup \{x'\}$ shatters $\mathcal{C}$. This is a contradiction because $\mathrm{VC}(\mathcal{C}) = d$. ■

D.3.1 Upper Bound on Sample Complexity.

Before proceeding with the proof, we clarify the notion of sample complexity used in this work. In the audit problem involving multiple populations, valid inference requires that every population contributes sufficient data. Let the set of populations be indexed by I, and for each $i \in I$, let m_i denote the minimum number of samples required from population i to ensure correctness of the audit. To enforce a uniform sampling scheme, where the same number of samples is drawn from each group, we define the (scalar) sample complexity as $m \coloneqq \max_{i \in I} m_i$. Thus, collecting m samples from each population guarantees that every group $i \in I$ receives at least its required number m_i of samples, independent of the index set I. Our approach to sample complexity differs fundamentally from classical methods. While traditional bounds typically assume a homogeneous population or ignore group structure, our group fairness audit framework explicitly accounts for data drawn from two distinct protected groups, requiring new complexity analyses to ensure equitable performance.

An important practical implication follows: Auditors may face strategic manipulations from the model owner that adjust the relative proportions between groups. This choice of sample complexity thereby manages the problem of class imbalance with considerable flexibility, provided that the sample size for each group never falls below this established threshold.

Let $\Big(\mathcal{X} \times \mathcal{Y}, \Omega(\mathcal{X} \times \mathcal{Y}), \mathcal{D}\Big)$ denote a probabilistic space, where $\mathcal{X}$ can be uncountable. Let $\mathcal{F}$ denote the strategic class, containing functions defined from $\mathcal{X}$ to $\mathcal{Y}$. And let $\mathcal{C}$ denote the corresponding concept class,

$$\mathcal{C} = \Big\{ c : c \subseteq \mathcal{X}, \exists \mathbb{1}_c \in \mathcal{F}, x \in c \iff \mathbb{1}_c(x) = 1 \Big\}$$

For a sample S of size $2m$, we use the notation $S^{\triangleright}$ to denote the first m instances in sample S, and $S^{\triangleleft}$ to denote the remaining m instances in sample S (i.e. $S = \langle S^{\triangleright}, S^{\triangleleft} \rangle$).

In the following, $\mathfrak{p}$ denotes the SP dimension of $\mathcal{F}$ (therefore of $\mathcal{C}$), for every protected group samples S_0 and S_1 of sizes m_0 and m_1 respectively, $\Pi^{\mathrm{SP}}_{\mathfrak{p}}(2m_0, 2m_1)$ denotes the SP-growth function defined as follows:

$$\Pi^{\mathrm{SP}}_{\mathfrak{p}}(m_0, m_1) = \max_{\substack{S : |S| = m_0 + m_1 \\ SP(\mathcal{F}) = \mathfrak{p}}} |\Delta^{SP}_{\mathcal{F}}(S_0, S_1)|$$

We first prove the following theorem:

Theorem 16. *Let $\epsilon, \alpha \in (0,1)$ and $\mathcal{C}$ a concept class of SP dimension $\mathfrak{p}$. The probability that the empirical SP-auditing risk of at least one concept differs from the true SP-auditing risk by more than ϵ in an i.i.d sample $S \sim \mathcal{D}$ of size $m_0 + m_1$ with $\min(m_0, m_1) \geq \frac{4}{\epsilon^2} \log \frac{1}{1-\alpha}$ satisfies the following inequality:*

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \Big[\sup_{c \in \mathcal{C}} |\mu_{\mathcal{D}}(c) - \mu_S(c)| > \epsilon \Big] \leq \alpha \Pi^{\mathrm{SP}}_{\mathfrak{p}}(2m_0, 2m_1) \exp \frac{-\min(m_0, m_1)}{16} \epsilon^2$$

As we will see in the proof, α can be any constant in $(0,1)$ depending on the problem parameters. By proving this result, the proof follows from Claim 22.

We begin by proving the following lemma:

Lemma 17. *For all $\alpha \in (0,1)$, $\min(m_0, m_1) \geq \frac{4}{\epsilon^2} \log \frac{1}{1-\alpha}$,*

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \Big[\sup_{c \in \mathcal{C}} |\mu_{\mathcal{D}}(c) - \mu_S(c)| > \epsilon \Big] \leq \alpha \mathop{\mathbb{P}}_{S \sim \mathcal{D}^{2m}} \Big[\sup_{c \in \mathcal{C}} |\mu_{S^{\triangleright}}(c) - \mu_{S^{\triangleleft}}(c)| > \frac{\epsilon}{2} \Big]$$

Interpretation of the result: For a sample size of $\frac{4}{\epsilon^2} \log \frac{1}{1-\alpha}$, if we repeat the experiment of measuring the random quantity of empirical statistical parity twice, the probability that the outcomes deviate by at least half of epsilon between the two experiments provides a lower bound on the probability that the empirical estimate deviates by more than epsilon from the true value of statistical parity. This provides the first step to prove uniform convergence.

Now we proceed with the proof of Lemma 17:

Proof

Let $\mathcal{K}_{\epsilon,2m}$ denote the event[5]:

$$\mathcal{K}_{\epsilon,2m} \triangleq \left\{ S : S \subseteq (\mathcal{X} \times \mathcal{Y})^{2m}, \sup_{c \in \mathcal{C}} |\mu_{S^{\triangleright}}(c) - \mu_{S^{\triangleleft}}(c)| > \frac{\epsilon}{2} \right\}$$

By the definition of expectation, we have,

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m}) = \int_{S \in (\mathcal{X} \times \mathcal{Y})^{2m}} \mathbb{1}_{\mathcal{K}_{\epsilon,m}}(S)\, d\mathcal{D}^{2m}(S)$$

By Fubini's theorem,

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m}) = \int_{S^{\triangleright} \in (\mathcal{X} \times \mathcal{Y})^{m}} \int_{S^{\triangleleft} \in (\mathcal{X} \times \mathcal{Y})^{m}} \mathbb{1}_{\mathcal{K}_{\epsilon,m}}(\langle S^{\triangleright}, S^{\triangleleft} \rangle)\, d\mathcal{D}^{m}(S^{\triangleright})\, d\mathcal{D}^{m}(S^{\triangleleft})$$

Let $\mathcal{J}_{\epsilon,m}$ denote the event we seek to establish an upper bound on its probability:

$$\mathcal{J}_{\epsilon,m} \triangleq \left\{ S : S \subseteq (\mathcal{X} \times \mathcal{Y})^{m}, \sup_{c \in \mathcal{C}} |\mu_{\mathcal{D}}(c) - \mu_{S}(c)| > \epsilon \right\}$$

Since $\mathcal{J}_{\epsilon,m} \subseteq (\mathcal{X} \times \mathcal{Y})^{m}$, we obtain:

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m}) \geq \int_{S^{\triangleright} \in \mathcal{J}_{\epsilon,m}} \int_{S^{\triangleleft} \in (\mathcal{X} \times \mathcal{Y})^{m}} \mathbb{1}_{\mathcal{K}_{\epsilon,m}}(\langle S^{\triangleright}, S^{\triangleleft} \rangle)\, d\mathcal{D}^{m}(S^{\triangleright})\, d\mathcal{D}^{m}(S^{\triangleleft}) \tag{2}$$

By definition of $\mathcal{J}_{\epsilon,m}$, for every $S^{\triangleright}$ in $\mathcal{J}_{\epsilon,m}$, there exists a concept $c_{S^{\triangleleft}}$ that verifies the following inequality:

$$|\mu_{\mathcal{D}}(c_{S^{\triangleright}}) - \mu_{S^{\triangleright}}(c_{S^{\triangleright}})| > \epsilon$$

On the other hand, by the inverse triangle inequality, we have for any $S^{\triangleright}$ in $\mathcal{J}_{\epsilon,m}$ and any $S^{\triangleleft}$ in $(\mathcal{X} \times \mathcal{Y})^{m}$:

$$|\mu_{S^{\triangleright}}(\mu_{S^{\triangleright}}) - \mu_{S^{\triangleleft}}(c_{S^{\triangleright}})| \geq |\mu_{S^{\triangleright}}(c_{S^{\triangleright}}) - \mu_{\mathcal{D}}(c_{S^{\triangleright}})| - |\mu_{S^{\triangleleft}}(c_{S^{\triangleright}}) - \mu_{\mathcal{D}}(c_{S^{\triangleright}})|$$

We deduce that for any $S^{\triangleleft}$ in $(\mathcal{X} \times \mathcal{Y})^{m}$ a sufficient condition to have $\langle S^{\triangleright}, S^{\triangleleft} \rangle \in \mathcal{K}_{\epsilon,2m}$ is $|\mu_{S^{\triangleleft}}(c_{S^{\triangleright}}) - \mu_{\mathcal{D}}(c_{S^{\triangleright}})| \leq \frac{\epsilon}{2}$. Let $\mathcal{A}_{S^{\triangleright}}$ denote the set of these events:

$$\mathcal{A}_{S^{\triangleright}} \triangleq \left\{ S^{\triangleleft} : S^{\triangleleft} \subseteq (\mathcal{X} \times \mathcal{Y})^{m}, |\mu_{S^{\triangleleft}}(c_{S^{\triangleright}}) - \mu_{\mathcal{D}}(c_{S^{\triangleright}})| \leq \frac{\epsilon}{2} \right\}$$

We have shown that:

[5] The event in the right side of the inequality in Lemma 17.

$$S^{\triangleleft} \in \mathcal{A}_{S^{\triangleright}} \implies \langle S^{\triangleright}, S^{\triangleleft} \rangle \in \mathcal{K}_{\epsilon,2m}$$

By implementing this in inequality 2, we obtain:

$$\underset{S\sim\mathcal{D}^{2m}}{\mathbb{P}}(\mathcal{K}_{\epsilon,2m}) \geq \int_{S^{\triangleright}\in\mathcal{J}_{\epsilon,m}} \int_{S^{\triangleleft}\in(\mathcal{X}\times\mathcal{Y})^m} \mathbb{1}_{\mathcal{A}_{S^{\triangleright}}}(S^{\triangleleft})\, d\mathcal{D}^m(S^{\triangleright})\, d\mathcal{D}^m(S^{\triangleleft}) \tag{3}$$

By using the result in Lemma 24, we have for all $S^{\triangleright}$ in $\mathcal{X}^m$:

$$\begin{aligned}\underset{S^{\triangleleft}\sim\mathcal{D}^m}{\mathbb{P}}\Big[|\mu_{S^{\triangleleft}}(c_{S^{\triangleright}}) - \mu_{\mathcal{D}}(c_{S^{\triangleright}})| \leq \frac{\epsilon}{2}\Big] &\geq 1 - \exp\Big\{\frac{-m_0 m_1}{2(m_0+m_1)}\epsilon^2\Big\} \\ &\geq 1 - \exp\Big\{-\frac{\min(m_0,m_1)\epsilon^2}{4}\Big\}\end{aligned}$$

Where the second step follows from Claim 21. The right side is bigger than α for $\min(m_0, m_1) \geq \frac{1}{\epsilon^2}\log\frac{1}{1-\alpha}$. We have shown that for any m such that, $\min(m_0, m_1) \geq \frac{4}{\epsilon^2}\log\frac{1}{1-\alpha}$, and any $S^{\triangleright} \in \mathcal{J}_{\epsilon,m}$:

$$\int_{S^{\triangleleft}\in(\mathcal{X}\times\mathcal{Y})^m} \mathbb{1}_{\mathcal{A}_{S^{\triangleright}}}(S^{\triangleleft})\, d\mathcal{D}^m(S^{\triangleleft}) \geq \alpha$$

By implementing this in inequality 3, we obtain the desired result.

Remark: Measurability constraints in the integration. We can observe that $\mathcal{J}_{\epsilon,m}$ and $\mathcal{K}_{\epsilon,2m}$ are measurable events when $\mathcal{C}$ is countable or finite. In the general case, even when $\mathcal{C}$ is measurable, it does not imply that the events inside the integral are measurable. A relaxed assumption of measurability is assuming $\mathcal{C}$ is indexed by a collection of Borel sets in an Euclidean space (Ben-David et al. 1990; Pollard 2012).

■

Lemma 18.

$$\underset{S\sim\mathcal{D}^{2m}}{\mathbb{P}}(\mathcal{K}_{\epsilon,2m}) \leq \Pi_{\mathfrak{p}}^{SP}(2m_0, 2m_1)\exp\frac{-\min(m_0,m_1)}{16}\epsilon^2$$

WHere $\mathfrak{p} = \mathrm{SP}(\mathcal{C})$

Proof Let Π_{m_0,m_1} the symmetric group defined on $[2m_0] \times [2m_1]$:

$$\begin{aligned}\Pi_{m_0,m_1} \triangleq \Big\{(\pi_0,\pi_1) \in \mathfrak{S}_{2m}^2 :&\forall (i,j) \in [2m_0]\times[2m_1], \\ &(\pi_0(i) = i \wedge \pi_0(m_0+i) = m_0+i) \vee (\pi_0(i) = m_0+i \wedge \pi_0(m_0+i) = i), \\ &(\pi_1(i) = i \wedge \pi_1(m_1+i) = m_1+i) \vee (\pi_1(i) = m_1+i \wedge \pi_1(m_1+i) = i)\Big\}\end{aligned}$$

We have, $|\Pi_{2m}| = 2^{m_0+m_1} = 2^m$.

By the i.i.d assumption on sampling from $\mathcal{D}$, and given the definition of the permutation set (where the permutation acts independently over each protected group), for any $\pi \in \Pi_{2m}$ such that $\pi = (\pi_0, \pi_1)$:

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{2m}}(\pi(S)) &= \prod_{i=1}^{2m}\mathop{\mathbb{P}}_{\mathcal{D}}\Big((x_{\pi_0(i)}, y_{\pi_1(i)})\Big)\\ &= \prod_{i=1}^{2m}\mathop{\mathbb{P}}_{\mathcal{D}}\Big((x_i, y_i)\Big)\\ &= \mathop{\mathbb{P}}_{\mathcal{D}^{2m}}(S)\end{aligned}$$

Hence, for every permutation $\pi \in \Pi_{2m}$, such that $\pi = (\pi_0, \pi_1)$:

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m}) &= \int_{S\in(\mathcal{X}\times\mathcal{Y})^{2m}} \mathbb{1}_{\mathcal{K}_{\epsilon,2m}}(\pi(S))\, d\mathcal{D}^{2m}(S)\\ &= \frac{1}{2^m}\sum_{\pi\in\Pi_{2m}}\int_{S\in(\mathcal{X}\times\mathcal{Y})^{2m}} \mathbb{1}_{\mathcal{K}_{\epsilon,2m}}(\pi(S))\, d\mathcal{D}^{2m}(S)\\ &= \int_{S\in(\mathcal{X}\times\mathcal{Y})^{2m}} \frac{\sum_{\pi\in\Pi_{2m}}\mathbb{1}_{\mathcal{K}_{\epsilon,2m}}(\pi(S))}{2^m}\, d\mathcal{D}^{2m}(S)\end{aligned}$$

For a fixed $S \subseteq (\mathcal{X}\times\mathcal{Y})^{2m}$, if we denote $(\pi_0, \pi_1) \to T_S((\pi_0, \pi_1))$ ($\pi \to T_S(\pi)$) the random variable, such that $T_S(\pi) = \mathbb{1}_{\mathcal{K}_{\epsilon,2m}}(\pi(S))$, the term inside the integral can be seen as the expectation over Π_{2m} with the uniform distribution of $T_S(\pi)$.

We have,

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m}) = \int_{S\in(\mathcal{X}\times\mathcal{Y})^{2m}} \frac{\sum_{\pi\in\Pi_{2m}} T_S(\pi)}{2^m}\, d\mathcal{D}^{2m}(S)$$

$$\begin{aligned}\mathop{\mathbb{E}}_{\pi\sim\mathcal{U}(\Pi_{2m})}\Big(T_S(\pi)\Big) &= \mathop{\mathbb{P}}_{\substack{\pi_1\sim\mathcal{U}(\Pi_{2m_1})\\ \pi_1\sim\mathcal{U}(\Pi_{2m_1})}}\Big[\exists c\in\mathcal{C} : |\frac{1}{m_0}\sum_{i=1}^{m_0}\mathbb{1}_{c(x_{\pi_0(i)})=1} - \frac{1}{m_1}\sum_{i=1}^{m_1}\mathbb{1}_{c(x_{\pi_1(i)})=1}\\ &\quad - \frac{1}{m_0}\sum_{i=1}^{m_0}\mathbb{1}_{c(x_{\pi_0(m_0+i)})=1} + \frac{1}{m_1}\sum_{i=1}^{m_1}\mathbb{1}_{c(x_{\pi_1(m_1+i)})=1}| \geq \frac{\epsilon}{2}\Big]\\ &= \mathop{\mathbb{P}}_{\substack{\pi_1\sim\mathcal{U}(\Pi_{2m_1})\\ \pi_1\sim\mathcal{U}(\Pi_{2m_1})}}\Big[\exists c\in\mathcal{C} : |\frac{1}{m_0}\Big(\sum_{i=1}^{m_0}\mathbb{1}_{c(x_{\pi_0(i)})=1} - \sum_{i=1}^{m_0}\mathbb{1}_{c(x_{\pi_0(m_0+i)})=1}\Big)\\ &\quad - \frac{1}{m_1}\Big(\sum_{i=1}^{m_1}\mathbb{1}_c(x_{\pi_1(i)}) = 1 - \sum_{i=1}^{m_1}\mathbb{1}_{c(x_{\pi_1(m_1+i)})=1}\Big)| \geq \frac{\epsilon}{2}\Big)\Big]\end{aligned}$$

For each fixed $c \in \mathcal{C}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\pi_0\sim\mathcal{U}(\Pi_{2m_0})}\Big[\Big(\sum_{i=1}^{m_0}\mathbb{1}_c(x_{\pi_0(i)}) = 1 - \sum_{i=1}^{m_0}\mathbb{1}_c(x_{\pi_0(m_0+i)}) = 1\Big)\Big] &= \frac{1}{2}\Big(\sum_{i=1}^{m_0}\mathbb{1}_c(x_i) = 1 - \sum_{i=1}^{m_0}\mathbb{1}_{\mathbb{1}_c(x_{m_0+i})=1}\Big) +\\ &\quad \frac{1}{2}\Big(\sum_{i=1}^{m_0}\mathbb{1}_c(x_{m_0+i}) = 1 - \sum_{i=1}^{m_0}\mathbb{1}_c(x_i) = 1\Big)\\ &= 0\end{aligned}$$

By symmetry of the permutations between the first and second protected groups, we have:

$$\mathop{\mathbb{E}}_{\pi_1\sim\mathcal{U}(\Pi_{2m_1})}\Big[\Big(\sum_{i=1}^{m_1}\mathbb{1}_c(x_{\pi_1(i)})=1-\sum_{i=1}^{m_1}\mathbb{1}_c(x_{\pi_1(m_1+i)})=1\Big)\Big]=0$$

Applying the Discrepancy Hoeffding inequality again (Lemma 24), for every $S\subseteq(\mathcal{X}\times\mathcal{Y})^{2m}$ and $c\in\mathcal{C}$,

$$\mathop{\mathbb{E}}_{\pi\sim\mathcal{U}(\Pi_{2m})}\Big(T_S(\pi)\Big)\leq\exp\frac{-\min(m_0,m_1)}{16}\epsilon^2$$

By applying the union bound over group-wise traces for the protected groups $\mathcal{X}_0$ and $\mathcal{X}_1$, exhibiting different discriminative behaviors, we have:

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{2m}}(\mathcal{K}_{\epsilon,2m})&\leq|\Delta^{\mathrm{SP}}_{\mathcal{C}}(S)|\exp\frac{-\min(m_0,m_1)}{16}\epsilon^2\\&\leq\Pi^{\mathrm{SP}}_{\mathfrak{p}}(2m_0,2m_1)\exp\frac{-\min(m_0,m_1)}{16}\epsilon^2\end{aligned}$$

The final step by applying Sauer lemma (Claim 25) twice on each protected group, which is true since an SP dichotomy exist when each of the points conditionned on each of the protected group define a dichotomy (can be classified following a concept c in $\mathcal{C}$).

■

D.4 Lower bounds on sample complexity: Hardness of weakly auditing

Let $(\epsilon,\delta)\in(0,1)^2$ and $\mathcal{C}$ denote a concept class of VC dimension $d+1$. Let $\mathcal{Z}=\{x_0,x_1\cdots,x_d\}$ be a subset of $\mathcal{X}$ that shatters $\mathcal{C}$. For any finite subset S of $\mathcal{X}$, we denote S_0 (resp. S_1) the subset of S whose elements belong to the first (resp. second) protected group.

For any $c,c'\in\mathcal{C}$, let $c\Delta^1_0c'=\{(x,x')\in\mathcal{X}_0\times\mathcal{X}_1:\mathbb{1}_{x\in c}-\mathbb{1}_{x'\in c}\neq\mathbb{1}_{x\in c'}-\mathbb{1}_{x'\in c'}\}$. Intuitively, this defines the set of pairs from the two protected groups, where c and c' behave differently [6]. For any concept c, we say that a concept c' is SP- consistent with (S,c) if for every pair $(x,x')\in S_0\times S_1$, $\mathbb{1}_{x\in c}-\mathbb{1}_{x'\in c}=\mathbb{1}_{x\in c'}-\mathbb{1}_{x'\in c'}$

By Lemma 3, there exists two points for $\mathcal{Z}$ such that each one belongs to a different protected group. Without loss of generality, we assume that $x_0\in\mathcal{X}_0$ and $x_1\in\mathcal{X}_1$. Let $\mathcal{Z}_0=\mathcal{Z}\cap\mathcal{X}_0$ and $\mathcal{Z}_1=\mathcal{Z}\cap\mathcal{X}_1$, and $d_0=|\mathcal{Z}_0|$, $d_1=|\mathcal{Z}_1|$, where $\mathcal{Z}_0$ (resp. $\mathcal{Z}_1$) is indexed by $\mathcal{I}_0$ (resp. $\mathcal{I}_1$).

Let $\mathcal{D}_{\mathcal{X}}$ denotes the marginal distribution on $\mathcal{X}$ supported on $\mathcal{Z}$, and defined as

$$\begin{cases}\mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{X}}}\{x_0\}=\frac{1-8\epsilon}{2}\\\mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{X}}}\{x_1\}=0\\\forall i\in\{2,3\cdots,d\}:\mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{X}}}\{x_i\}=\frac{8\epsilon}{d-1}\end{cases}$$

Since $\mathcal{D}_{\mathcal{X}}$ is supported on $\mathcal{Z}$, we assume without loss of generality that $\mathcal{X}=\mathcal{Z}$ and $\mathcal{C}\subseteq 2^{\mathcal{Z}}$.

Let $\mathcal{C}_{0,1}=\Big\{\{x_0,x_1\}\cup T,T\subseteq\{x_2,x_3,\cdots,x_d\}\Big\}$

And $\tilde{S}_m=\{S\subseteq\mathcal{X}:|S_0|=m_0,|S_1|=m_1,m_0\leq\frac{d_0}{2},m_1\leq\frac{d_1}{2}\}$

[6] c and c' do not behave the same with respect to the pair (x,x').

We further assume that the auditor $\mathbb{A}$ is (possibly) randomized, and takes as input ω denoting a sequence of boolean random variables sampled independently (i.e. $\mathbb{A} = \mathbb{A}(S;\omega)$). And without loss of generality, we assume that $(x_0, x_1) \in \mathbb{A}(S;\omega)$ whenever concepts are selected from $\mathcal{C}_{0,1}$.

For any fixed $c \in \mathcal{C}$ and $S \in \tilde{S}_m$, let $p = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{A}(S;\omega)\Big]$ and $p' = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[c\Big]$

$$\begin{aligned}
|\mu(\mathbb{A}(S;\omega)) - \mu(c)| &= |\sum_{\substack{k\in\mathcal{I}_0\\k\neq 0}}\sum_{x_k\in\mathbb{A}(S;\omega)} p_k - \sum_{\substack{k\in\mathcal{I}_1\\k\neq 1}}\sum_{x_k\in\mathbb{A}(S;\omega)} p_k - \sum_{\substack{k\in\mathcal{I}_0\\k\neq 0}}\sum_{x_k\in c} p'_k + \sum_{\substack{k\in\mathcal{I}_1\\k\neq 1}}\sum_{x_k\in c} p'_k| \\
&= \frac{8\epsilon}{d-1}|\sum_{\substack{k\in\mathcal{I}_0\\k\neq 0}}\sum_{x_k\in\mathbb{A}(S;\omega)} 1 - \sum_{\substack{k\in\mathcal{I}_1\\k\neq 1}}\sum_{x_k\in\mathbb{A}(S;\omega)} 1 - \sum_{\substack{k\in\mathcal{I}_0\\k\neq 0}}\sum_{x_k\in c} 1 + \sum_{\substack{k\in\mathcal{I}_1\\k\neq 1}}\sum_{x_k\in c} 1| \\
&= \frac{8\epsilon}{d-1}\Big|(x,x') \in \mathcal{Z}_0\times\mathcal{Z}_1 : (x,x') \in \mathbb{A}(S;\omega)\Delta_0^1 c\Big|
\end{aligned}$$

On the other hand, suppose we have a uniform distribution over the concept class $\mathcal{C}_{0,1}$. For each $c \in \mathcal{C}_{0,1}$, there are exactly $2^{d_0-m_0}$ (resp. $2^{d_1-m_1}$) concepts that are consistent with (S_0, c) (resp. (S_1, c)). For any of the couples $(x,x') \in \mathcal{Z}_0\times\mathcal{Z}_1$ that are not in S, $\frac{1}{2}$ of the SP-consistent concepts will contain this couple and $\frac{1}{2}$ will not. We deduce that $\mathbb{A}(S;\omega)$ behaves the same as c with respect to this couple for exactly $\frac{1}{2}$ of these $2^{d_0-m_0+d_1-m_1}$ SP-consistent concepts. Therefore:

$$\begin{aligned}
\mathop{\mathbb{E}}_{c\sim\mathbb{U}_{\mathcal{C}_{0,1}}}\Big[\Big|(x,x') \in \mathcal{Z}_0\times\mathcal{Z}_1 : (x,x') \in \mathbb{A}(S;\omega)\Delta_0^1 c\Big|\Big] &\geq \frac{d_0 - m_0 + d_1 - m_1}{2} \\
&= \frac{d-m}{2} \\
&\geq \frac{d}{4}
\end{aligned}$$

Where the second step follows from the fact that $S \in \tilde{S}_m$. Since $d = d_0 + d_1$, we have shown that:

$$\mathop{\mathbb{E}}_{c\sim\mathbb{U}_{\mathcal{C}_{0,1}}}\Big[|\mu(\mathbb{A}(S;\omega)) - \mu(c)|\Big] \geq \frac{2\epsilon d}{d-1} \geq 2\epsilon \tag{4}$$

This is true for every value of S and ω,

$$\mathop{\mathbb{E}}_{c\sim\mathbb{U}_{\mathcal{C}_{0,1}},S,\omega}\Big[|\mu(\mathbb{A}(S;\omega)) - \mu(c)|\Big] \geq \frac{2\epsilon d}{d-1} \geq 2\epsilon$$

Therefore there exists $c \in \mathcal{C}_{0,1}$, such that:

$$\mathop{\mathbb{E}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega)) - \mu(c)|\Big] \geq \frac{2\epsilon d}{d-1} \geq 2\epsilon$$

On the other hand, we have for all $(x,x') \in \mathcal{X}_0\times\mathcal{X}_1$:

$$\Big(\mathbb{1}_{x\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c}\Big) \vee \Big(\mathbb{1}_{x\in\mathbb{A}(S;\omega)} \neq \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\in\mathbb{A}(S;\omega)} \neq \mathbb{1}_{x'\in c}\Big) \iff (x,x') \notin \mathbb{A}(S;\omega)\Delta_0^1 c$$

We deduce,

$$
\begin{aligned}
1 - \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{A}(S;\omega)\Delta_0^1 c | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] &= \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c}\Big) \\
&\quad \vee \Big(\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\leq \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\in\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\quad + \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\leq 1 - \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \vee \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\quad + \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big]
\end{aligned}
$$

$$
\begin{aligned}
\mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{A}(S;\omega)\Delta_0^1 c | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] &\geq \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \vee \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\quad - \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} \wedge \mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \\
&\geq \mathop{\mathbb{P}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} | x\in\mathcal{X}_0\Big] + \mathop{\mathbb{P}}_{x'\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x'\in\mathcal{X}_1\Big]
\end{aligned}
$$

Lemma 19.

$$
\mathop{\mathbb{P}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} | x\in\mathcal{X}_0\Big] + \mathop{\mathbb{P}}_{x'\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x'\in\mathcal{X}_1\Big] \geq |\mu(\mathbb{A}(S;\omega)) - \mu(c)|
$$

Proof By triangle inequality,

$$
\begin{aligned}
|\mu(\mathbb{A}(S;\omega)) - \mu(c)| &\leq |\mu^0(\mathbb{A}(S;\omega)) - \mu^0(c)| + |\mu^1(\mathbb{A}(S;\omega)) - \mu^1(c)| \\
&\leq |\mu^0(\mathbb{A}(S;\omega)) - \mu^0(c)| + |\mu^1(\mathbb{A}(S;\omega)) - \mu^1(c)| \\
&\leq \mathop{\mathbb{P}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x\in c} | x\in\mathcal{X}_0\Big] + \mathop{\mathbb{P}}_{x'\sim\mathcal{D}_\mathcal{X}}\Big[\mathbb{1}_{x'\notin\mathbb{A}(S;\omega)} = \mathbb{1}_{x'\in c} | x'\in\mathcal{X}_1\Big]
\end{aligned}
$$

■

We have proved the following:

$$
|\mu(\mathbb{A}(S;\omega)) - \mu(c)| \leq \mathop{\mathbb{P}}_{\substack{x\sim\mathcal{D}_\mathcal{X}\\x'\sim\mathcal{D}_\mathcal{X}}}\Big[\mathbb{A}(S;\omega)\Delta_0^1 c | x\in\mathcal{X}_0, x'\in\mathcal{X}_1\Big] \tag{5}
$$

On the other hand, since $\mathbb{A}(S;\omega)$ is always correct on (x_0, x_1) with respect to c, then for all $(x, x')\in\mathcal{X}_0\times\mathcal{X}_1$:

$$
(x, x')\in\mathbb{A}(S;\omega)\Delta_0^1 c \implies (x, x')\neq(x_0, x_1)
$$

We deduce:

$$\begin{aligned}\mathop{\mathbb{P}}_{\substack{X\sim\mathcal{D}_{\mathcal{X}}\\X'\sim\mathcal{D}_{\mathcal{X}}}}\Big[\mathbb{A}(S;\omega)\Delta_0^1 c|X\in\mathcal{X}_0,X'\in\mathcal{X}_1\Big] &\leq \mathop{\mathbb{P}}_{\substack{X\sim\mathcal{D}_{\mathcal{X}}\\X'\sim\mathcal{D}_{\mathcal{X}}}}\Big[X\neq x_0\wedge X'\neq x_1|X\in\mathcal{X}_0,X'\in\mathcal{X}_1\Big]\\ &\leq \mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\mathcal{X}}}\Big[X\neq x_0]\\ &= 8\epsilon\end{aligned}$$

Implementing this in the inequality 5 gives:

$$|\mu(\mathbb{A}(S;\omega))-\mu(c)|\leq 8\epsilon \tag{6}$$

From the inequalities in 4 and 6, we get:

$$2\epsilon\leq \mathop{\mathbb{E}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|\Big]\leq 8\epsilon\mathop{\mathbb{P}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|>\epsilon\Big]+\epsilon\Big(\mathop{\mathbb{P}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|>\epsilon\Big]-1\Big)$$

Therefore,

$$\mathop{\mathbb{P}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|>\epsilon|S\in\tilde{S}_m\Big]>\frac{1}{7}$$

On the other hand,

$$\begin{aligned}\mathop{\mathbb{P}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|>\epsilon\Big] &= \mathop{\mathbb{P}}_{S,\omega}\Big[|\mu(\mathbb{A}(S;\omega))-\mu(c)|>\epsilon|S\in\tilde{S}_m\Big]\mathop{\mathbb{P}}_{S}\Big[S\in\tilde{S}_m\Big]\\ &\geq \frac{1}{7}\mathop{\mathbb{P}}_{S}\Big[S\in\tilde{S}_m\Big]\end{aligned}$$

Let T_0 (resp. T_1) denote the number of realizations of $\{x_2,x_3\ldots,x_d\}$ in a sample S_0 (resp. S_1)

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{D}^m}\Big\{\tilde{S}_m\Big\} &\geq \mathop{\mathbb{P}}_{\mathcal{D}^m}\Big\{T_0\leq\frac{d_0}{2}\wedge T_1\leq\frac{d_1}{2}\Big\}\\ &= \mathop{\mathbb{P}}_{\mathcal{D}^m}\Big\{\min(T_0,T_1)\leq\min(\frac{d_0}{2},\frac{d_1}{2})\Big\}\end{aligned}$$

The last term is bounded by the probability that a binomial($\min(m_0,m_1),8\epsilon$) is less than $\min(\frac{d_0}{2},\frac{d_1}{2})$.
By applying Bernstein inequality 23,

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{D}^m}\Big\{\tilde{S}_m\Big\} &\geq 1-\exp-\frac{\min(\frac{d_0}{2},\frac{d_1}{2})-8\min(m_0,m_1)\epsilon^2}{8\min(m_0,m_1)\epsilon(1-8\epsilon)}\\ &\geq 1-\exp-\frac{\min(d_0,d_1)-16\min(m_0,m_1)\epsilon^2}{16\min(m_0,m_1)\epsilon(1-8\epsilon)}\end{aligned}$$

For $\min(m_0,m_1)=\frac{\min(d_0,d_1)}{32\epsilon}$, we can simplify the right term and for $\epsilon<\frac{1}{8},\delta<\frac{1}{20}$ the sample complexity $\min(m_0,m_1)>\frac{\min(d_0,d_1)}{32\epsilon}$.

E Proofs for strongly auditable classes

E.1 All finite classes are strongly auditable (Theorem 7)

We start by restating Theorem 7:

Theorem 7 (Strongly auditable finite classes)**.** *Any finite hypothesis class is strongly auditable with a sample complexity* $\mathcal{O}\Big(\max\Big\{\underbrace{\frac{1}{\epsilon^2}\log\frac{|\mathcal{F}|}{\delta}}_{\text{without prospect}}, \underbrace{\frac{1}{\log\frac{1}{\epsilon^2}}\log\frac{|\mathcal{F}|}{\delta}}_{\text{with prospect}}\Big\}\Big)$.

Proof Let m^{corr} (resp. m^{comp}) denote the sample complexity for correctness (resp. completeness). Let f^* denote the black-box model under audit accessible through a set of labeled samples S, $S_{|x}$ denotes the projection of S on the input space $\mathcal{X}$. μ^* denotes the true value of statistical parity of the model under audit f^*, i.e., $\mu^* = \mu(f^*)$

Step 1: Bounding sample complexity for correctness m^{corr}. Let $\tilde{\mathcal{F}}$ denote the set $\{f \in \mathcal{F}, |\mu_{\mathcal{D}}(f) - \mu^*| > \epsilon\}$. In the following, we decouple the subclass $A[S]$ from S using the set $\tilde{\mathcal{F}}$ by showing the following:

Since the realizability assumption holds, and $A[S] = \{f \in \mathcal{F}, \mu_S(f) = \mu_S(f^*)\}$, we have: $\{S_{|x}, \exists h \in A[S], |\mu_{\mathcal{D}}(f) - \mu^*| > \epsilon\} \subseteq \{S_{|x}, \exists h \in \tilde{\mathcal{F}}, \mu_S(f) = \mu_S(f^*)\}$

We deduce,

$$
\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{corr}}}}\Big[\exists h \in A[S], |\mu_{\mathcal{D}}(f) - \mu^*| > \epsilon\Big] &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{corr}}}}\Big[\exists h \in \tilde{\mathcal{F}}, \mu_S(f) = \mu_S(f^*)\Big] \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{corr}}}}\Big[\mu_S(f) = \mu_S(f^*)\Big] \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{corr}}}}\Big[\forall (x_0, x_1) \in S_0 \times S_1 : (f(x_0) = y_0 \wedge f(x_1) = y_1) \vee (f(x_0) \neq y_0 \wedge f(x_1) \neq y_1)\Big] \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \prod_{i=1}^{m^{\text{corr}}} \mathop{\mathbb{P}}_{x_i\sim\mathcal{D}}\Big[h(x_i) = y_i\Big] \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \prod_{i=1}^{m^{\text{corr}}} (1-\epsilon) \\
&\leq |\mathcal{F}| e^{-\epsilon m^{\text{corr}}}
\end{aligned}
$$

The first inequality comes from the result of the claim, the second inequality comes from the union-bound, the fourth inequality comes from the i.i.d assumption, and the fifth inequality comes from the definition of $\tilde{\mathcal{F}}$.

This result shows that a sample size of $\mathcal{O}\Big(\frac{1}{\epsilon}\log\frac{|\mathcal{F}|}{\delta}\Big)$ is sufficient for correctness.

Step 2: Bounding sample complexity for completeness m^{comp}. In the following proof, we use the same decoupling argument with an extra cost over ϵ proportion of samples that leads to a small true error but a "large" (nonzero) empirical error. Let $\bar{\tilde{\mathcal{F}}}$ denote the complement of the set $\tilde{\mathcal{F}}$ defined in the previous step. That is $\bar{\tilde{\mathcal{F}}} = \{f \in \mathcal{F}, |\mu_{\mathcal{D}}(f) - \mu^*| \leq \epsilon\}$, and set $\mathcal{T}$ to be the set $T \triangleq \{S_{|x}, \exists f \in \bar{\tilde{\mathcal{F}}} \setminus A[S], \forall x \in S_{|x} : f(x) \neq y\}$.

Claim 20. *The following is true:*

$\{S_{|x}, \exists h \in \mathcal{F} \setminus A[S], |\mu_{\mathcal{D}}(f) - \mu^*| \leq \epsilon\} \subseteq \{S_{|x}, \exists h \in \bar{\tilde{\mathcal{F}}}, \mu_S(f) - \mu_S(f^*) = 0\} \cup \mathcal{T}$.

Proof: Let $S_{|x}$ and $h \in \mathcal{F} \setminus A[S]$ such that, $|\mu_{\mathcal{D}}(f) - \mu^*| \leq \epsilon$

Since $f \notin A[S]$, $\mu_{\mathcal{D}}(f) \neq \mu_S(f^*)$, but since $|\mu_{\mathcal{D}}(f) - \mu^*| \leq \epsilon$, let $\mathcal{T}$ denote the fraction of samples from $S_{|x}$ that induce the error, that is for all $x \in \mathcal{T}, h(x) \neq y$.

We have partitioned $S_{|x}$ into a subset S_1 that agrees with the true function on the labels, that is $\mu_{S_1}(f) = \mu_{S_1}(f^*)$, and another subset that verifies: $\mathop{\mathbb{P}}_{x\sim\mathcal{D}}(\mathcal{T}) \leq \epsilon$, which concludes the proof of the claim.

We deduce the following:

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{comp}}}}\left[\exists f \in \mathcal{F} \setminus A[S], |\mu_{\mathcal{D}}(f) - \mu^*| \leq \epsilon\right] &\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{comp}}}}\left[\exists f \in \bar{\bar{\mathcal{F}}}, \mu_S(f) = \mu_S(f^*)\right] + \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{comp}}}}\left[\forall x \in S, x \in \mathcal{T}\right] \\
&\leq \sum_{f\in\mathcal{F}} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{m^{\text{corr}}}}\left[\mu_S(f) = \mu_S(f^*)\right] + \sum_{f\in\mathcal{F}} \prod_{i=1}^{m^{\text{corr}}} \mathop{\mathbb{P}}_{x_i\sim\mathcal{D}}\left[\mathcal{T}\right] \\
&\leq \sum_{f\in\mathcal{F}} \prod_{i=1}^{m^{\text{corr}}} (1-\epsilon) + \sum_{h\in\mathcal{F}} \prod_{i=1}^{m^{\text{corr}}} \epsilon \\
&\leq |\mathcal{F}| e^{-\epsilon m^{\text{corr}}} + |\mathcal{F}| \epsilon^{m^{\text{corr}}}
\end{aligned}$$

This result shows that a sample size of $\mathcal{O}\Big(\max\Big\{\frac{1}{\epsilon}\log\frac{|\mathcal{F}|}{\delta}, \frac{1}{\log\frac{1}{\epsilon}}\log\frac{|\mathcal{F}|}{\delta}\Big\}\Big)$ is sufficient for correctness.

■

E.2 Concentration bounds on prospect ratio (Theorem 8)

We first begin by restating the theorem:

Theorem 8 (Concentration of prospect ratio). *For all* $\epsilon, \upsilon, \tau \in (0,1)$, $\mathbb{P}\{r(\epsilon-\upsilon) - \tau \leq \hat{r}_{n,m_0,m_1}(\epsilon) \leq r(\epsilon+\upsilon) + \tau\} \geq \Big(1 - \exp\Big\{\frac{-2\upsilon^2 m_0 m_1}{n(m_0+m_1)}\Big\}\Big)^n (1 - \exp(-n\tau^2))^2$.

Proof Let $\epsilon, \upsilon, \tau \in (0,1)$ By the definition of the estimator of prospect ratio:

$$\hat{\tilde{r}}_{n,m_0,m_1}(\epsilon) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{|\frac{1}{m_0}\sum_{i=1}^{m_0}\mathbb{1}_{f_k(x_i)=1} - \frac{1}{m_1}\sum_{j=1}^{m_1}\mathbb{1}_{f_k(x'_j)=1}| \leq \epsilon}$$

Since we sample independently f_k's from $\mathcal{F}$, the random variables $\mathbb{1}_{|\frac{1}{m_0}\sum_{i=1}^{m_0}\mathbb{1}_{f_k(x_i)=1} - \frac{1}{m_1}\sum_{j=1}^{m_1}\mathbb{1}_{f_k(x'_j)=1}| \leq \epsilon}$ are also independent and take values in $[0,1]$.

By Hoeffding inequality:

$$\mathbb{P}\{|\tilde{r}_n(\epsilon) - r(\epsilon)| \geq \tau\} \leq 2\exp\{-2n\tau^2\} \tag{7}$$

On the other hand, let S be sample that contains m_0 points from first protected group and m_1 points from second protected group. By applying Lemma 24 on each function f_k over equally size subsamples of size $\frac{m}{n}$, we have for all $k \in [n]$:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\frac{m}{n}}}\{|\hat{\mu}_S(f_k) - \mu_{\mathcal{D}}(f_k)| \geq \upsilon\} \leq \exp\Big\{\frac{-2\upsilon^2 m_0 m_1}{n(m_0+m_1)}\Big\}$$

By the independence of the events:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\frac{m}{n}}}\{\inf_{k\in[N]} |\hat{\mu}_S(f_k) - \mu_{\mathcal{D}}(f_k)| \geq \upsilon\} = \prod_{k=1}^{n} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^{\frac{m}{n}}}\{|\hat{\mu}_S(f_k) - \mu_{\mathcal{D}}(f_k)| \geq \upsilon\} \tag{8}$$

We deduce:

$$\mathbb{P}_{S\sim\mathcal{D}^{\frac{m}{n}}}\{\sup_{k\in[N]}|\hat{\mu}_S(f_k)-\mu_{\mathcal{D}}(f_k)|\leq\upsilon\}\geq\left(1-\exp\left\{\frac{-2\upsilon^2m_0m_1}{n(m_0+m_1)}\right\}\right)^n$$

Now let $k\in[N]$, such that $|\hat{\mu}_S(f_k)-\mu_{\mathcal{D}}(f_k)|\geq\upsilon$

for all τ in $(0,1)$,

$$\begin{aligned}f_k\in\hat{\mathcal{P}}(\mathcal{F},\epsilon)&\implies|\hat{\mu}-\mu(f_k)|\leq\epsilon\leq\epsilon+\upsilon\\&\implies f_k\in\hat{\mathcal{P}}(\mathcal{F},\epsilon+\upsilon)\end{aligned}$$

Since this is true for all k's, we deduce: $\hat{\tilde{r}}\leq\tilde{r}(\epsilon+\upsilon)$

Similarly if $f_k\in\hat{\mathcal{P}}(\mathcal{F},\epsilon-\upsilon)$, we have $\tilde{r}(\epsilon-\upsilon)\leq\hat{\tilde{r}}(\epsilon)$ In other words, $\tilde{r}(\epsilon-\upsilon)\leq\hat{\tilde{r}}(\epsilon)\leq\tilde{r}(\epsilon+\upsilon)$

Therefore,

$$\mathbb{P}_{S\sim\mathcal{D}^{\frac{m}{n}}}\{\sup_{k\in[N]}|\hat{\mu}_S(f_k)-\mu_{\mathcal{D}}(f_k)|\leq\upsilon\}\leq\mathbb{P}\{\tilde{r}(\epsilon-\upsilon)\leq\hat{\tilde{r}}(\epsilon)\leq\tilde{r}(\epsilon+\upsilon)\}\tag{9}$$

From Equation 8 and Equation9, we deduce:

$$\left(1-\exp\left\{\frac{-2\upsilon^2m_0m_1}{n(m_0+m_1)}\right\}\right)^n\leq\mathbb{P}\{\tilde{r}(\epsilon-\upsilon)\leq\hat{\tilde{r}}(\epsilon)\leq\tilde{r}(\epsilon+\upsilon)\}\tag{10}$$

By the inequality in 7 and the inequality 10, we deduce the desired result.

■

E.3 Infinite classes are not auditable (Proposition 9)

We start be restating Proposition 9:

Proposition 9. *Any class of infinite VC dimension can not be weakly or strongly auditable.*

Proof We prove the result by showing a lower bound on the SP dimension that depends on the VC dimension.

Let $\mathcal{F}$ denotes a hypothesis class and S denotes a sample that SP-shatters $\mathcal{F}$, S_0 (resp. S_1) the subset of S belonging to the first protected group (resp. second protected group).

Since S SP-shatters $\mathcal{F}$, by the result in Lemma 3, we have :

$$\begin{aligned}\max(|S_0|,|S_1|)&\leq|S|-2\\2^{\max(|S_0|,|S_1|)}&\leq2^{|S|}(1-\frac{3}{4})\end{aligned}$$

Hence, for all S that SP-shatters $\mathcal{F}$:

$$2^{|S|}-2^{\max(|S_0|,|S_1|)}\geq\frac{3}{4}2^{|S|}$$

And since

$$2^{|S|}-2^{\max(|S_0|,|S_1|)}\leq2^{|S|}-2^{|S_0|}-2^{|S_1|}$$

We deduce:

$$\mathrm{SP}(\mathcal{F}) \geq \log_2 \frac{3}{4} + \mathrm{VC}(\mathcal{F})$$

This implies that if the VC dimension is infinite then SP dimension is also infinite. ■

F Extended Technical Details For Statistical Parity Audit

F.1 Hardness of Identifying Prospect Class.

When the prospect class is infinite, any algorithm attempting to exhaustively evaluate all models in the prospect class would never terminate. This challenge can be illustrated through linear classifiers in two dimensions, as shown in Figure 4. In this example, any line passing through the blue region belongs to the prospect class, resulting in infinitely many candidate models.

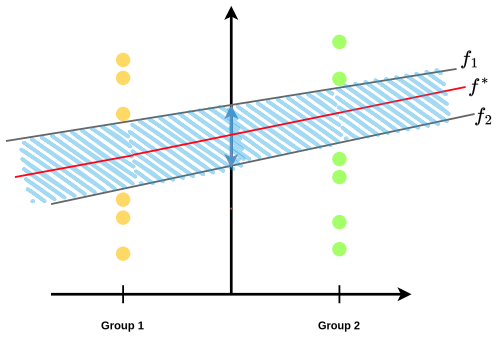


Figure 4: Illustration of prospect class for linear classifiers. Group 1 (yellow) and Group 2 (green) represent the two protected groups. Any classifier in the region delimited by f_1 and f_2 has the same statistical parity value.

F.2 Relationship between Prospect Ratio and SP-dimension.

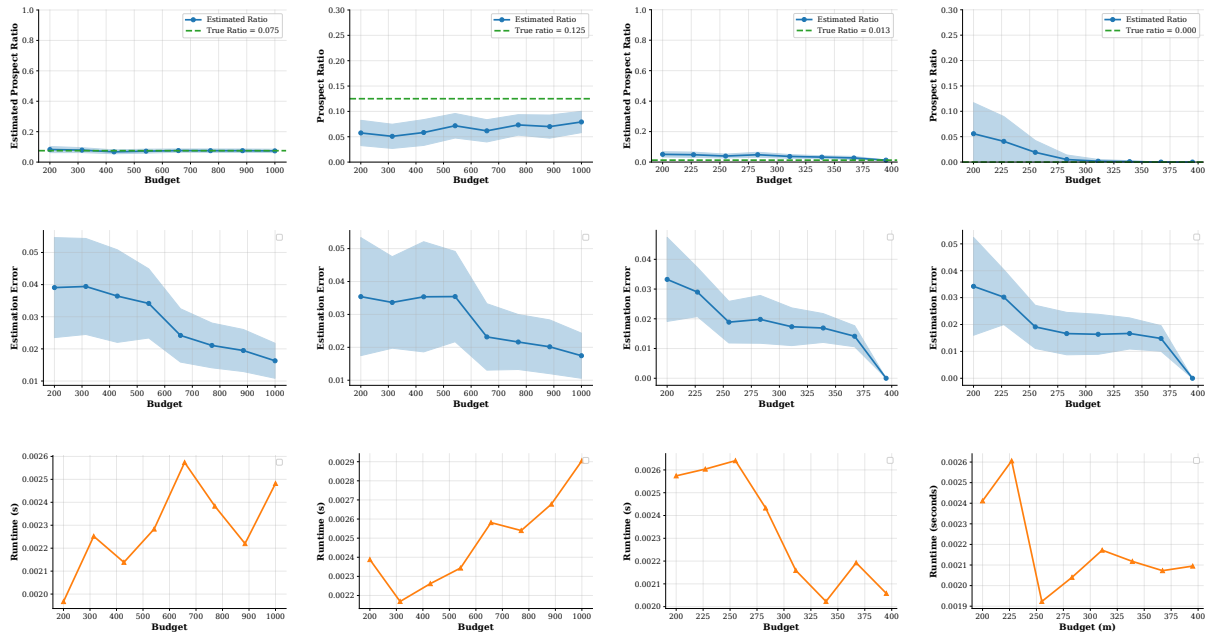
The prospect ratio and SP dimension serve fundamentally different purposes in our analysis. While the SP dimension, like other complexity measures such as VC dimension, characterizes the capacity of a hypothesis class to generalize finite-sample properties to distributional ones, the prospect ratio serves a distinct role. Specifically, it quantifies the likelihood of finding models that satisfy post-audit requirements while maintaining equivalent properties under audit values. A key distinction is that the prospect ratio is inherently data dependent, whereas the SP dimension is determined uniquely by the hypothesis class structure.

G Extended experimental details and additional datasets

G.1 Experiment & Implementation Details

All experiments were conducted on an 11th Gen Intel® Core™ i7-1185G7 processor (3.00 GHz, 8 cores) with 32.0 GiB of RAM. Implementation details and instructions for reproducing our results are provided in the supplementary code repository: https://anonymous.4open.science/r/Auditors-with-prospects-050F.

Figure 5 extends Figure 3 by reporting runtimes across datasets and strategic set transitions. In Experiment (a), the true prospect ratio (defined as the cardinality of the true prospect set divided by that of the sampled model from the strategic class) is 0.075. Despite this sparsity, Algorithm 1 accurately identifies the prospect set using only 200 samples. Notably, estimation accuracy improves monotonically with sample size, while runtime remains stable, exhibiting only bounded, non-monotonic fluctuations. In Experiment (b), where the true prospect ratio increases to 0.125, the algorithm similarly recovers the prospect subset with high fidelity, achieving vanishing statistical error and low computational overhead. We further validate our approach on the Student Performance



(a) COMPAS dataset MLPs as strategic set

(b) COMPAS dataset RFs as strategic set

(c) Student dataset MLPs as strategic set

(d) Student dataset RFs as strategic set

Figure 5: Comparison of Runtimes, errors in statistical parity estimation, and prospect ratio across different datasets and strategic sets.

Table 1: Summary of experimental results for a budget of 1000 samples.

Dataset	**Strategic Class**	**Estimation Error**	**Ratio Error**	**Runtime (ms)**
COMPAS	MLP	2.33×10^{-2}	2.5×10^{-3}	1.9
COMPAS	Random Forest	2.33×10^{-2}	5×10^{-3}	1.3
Student Performance	MLP	8.22×10^{-3}	0	1.7
Student Performance	Random Forest	8.22×10^{-3}	2.77×10^{-17}	1.2

dataset (Silva 2008), using gender as the protected attribute, and observe consistent behavior: the auditor reliably captures the full prospect class with minimal runtime and negligible estimation error.

These findings align with the broader empirical analysis in Section 5. As shown in Figures 3 and 5, the auditor effectively recovers the entire prospective class, as measured by the prospect ratio. Concurrently, it preserves statistical parity, demonstrating stability of fairness properties under audit. Crucially, this is achieved without sacrificing flexibility in model updates and predictive performance, illustrating a favorable trade-off between accuracy and completeness.

Applicability to LLM Auditing. While our experiments focus on traditional predictive models, our framework is particularly well-suited for auditing dynamic systems such as large language models (LLMs). LLMs undergo frequent updates that can shift their fairness behavior over time (Chen et al. 2024; Schaeffer et al. 2023), and recent benchmarks reveal substantial subgroup disparities across models (Parrish et al. 2021). Our estimator's sample efficiency, stability under distribution shift, and ability to operate without full model access make it a promising candidate for continuous fairness monitoring in such settings, a direction we leave for future work.

H Useful Technical Results

Claim 21. *For all $m_0, m_1 \in \mathbb{R}^+$,*

$$\min(m_0, m_1) \leq \frac{2m_0 m_1}{m_0 + m_1}$$

Claim 22. *For $a \geq 1, b > 0$, if the following holds:*

$$x \geq 4a \log 2a + 2b$$

We have:

$$x \geq a \log x + b$$

Claim 23 (Bernstein inequality)**.** *When $X_1, \cdots, X_n$ are independent random variables, with $\mathbb{P}[X_i] \leq M$ almost surely for all $i \in [n]$, the following holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sum_{i=1}^{n}\mathbb{E}[X_i^2] + Mt/3}\right)$$

Lemma 24 (Discrepancy Chernoff bounds)**.** *If $Q_1, \cdots, Q_{m_0}, R_1, \cdots, R_{m_1}$ are independent random variables taking values in $[0, 1]$, then:*

$$\mathbb{P}\left[Q_{m_0} - R_{m_1} - (\mathbb{E}Q_{m_0} - \mathbb{E}R_{m_1}) > \epsilon\right] \leq \exp\frac{-2m_0 m_1 \epsilon^2}{m_0 + m_1}$$

The following claim will serve to derive upper bounds for improper auditing.

Claim 25. *For all $m, s \in \mathbb{N}$,*

$$\sum_{i=0}^{s}\binom{m}{i} \leq \left(\frac{en}{s}\right)^s$$