Distribution-Specific Auditing for Subgroup Fairness

Daniel Hsu
Columbia University, New York, NY, USA

Jizhou Huang
Washington University in St. Louis, MO, USA

Brendan Juba
Washington University in St. Louis, MO, USA

Abstract

We study the problem of auditing classifiers for statistical subgroup fairness. Kearns et al. [20] showed that the problem of auditing combinatorial subgroups fairness is as hard as agnostic learning. Essentially all work on remedying statistical measures of discrimination against subgroups assumes access to an oracle for this problem, despite the fact that no efficient algorithms are known for it. If we assume the data distribution is Gaussian, or even merely log-concave, then a recent line of work has discovered efficient agnostic learning algorithms for halfspaces. Unfortunately, the reduction of Kearns et al. was formulated in terms of weak, "distribution-free" learning, and thus did not establish a connection for families such as log-concave distributions. In this work, we give positive and negative results on auditing for Gaussian distributions: On the positive side, we present an alternative approach to leverage these advances in agnostic learning and thereby obtain the first polynomial-time approximation scheme (PTAS) for auditing nontrivial combinatorial subgroup fairness: we show how to audit statistical notions of fairness over homogeneous halfspace subgroups when the features are Gaussian. On the negative side, we find that under cryptographic assumptions, no polynomial-time algorithm can guarantee any nontrivial auditing, even under Gaussian feature distributions, for general halfspace subgroups.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Fairness auditing, agnostic learning, intractability

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.5

Related Version *Previous Version*: https://arxiv.org/abs/2401.16439

Funding This work was partially supported by NSF awards IIS-2040971, IIS-1908287, and IIS-1942336, and NSF-Amazon award IIS-1939677.

1 Introduction

The deployment of decision rules obtained using machine learning has raised the risk that the rules may exhibit biases against historically marginalized communities. In particular, Kearns et al. [20] raised the concern that these decision rules may be biased against sub-groups characterized by a combination of "protected" attributes. Since there are an exponential number of such subgroups, even detecting such statistical patterns of discrimination is a nontrivial computational problem; indeed, Kearns et al. [20] showed that the problem of finding disadvantaged subgroups is equivalent to the problem of agnostic learning, which is believed to be intractable in general for all but the simplest classes of sets. Essentially all work [20, 23, 18] on remedying statistical measures of discrimination against subgroups assumes access to an oracle for this problem, despite the fact that no efficient algorithms are known for it. In this work we are proposing a solution for a variant of the fairness

auditing problem with provable guarantees of efficiency and correctness, as well as some strong limitations on the extent to which these solutions can be extended to richer families of subgroups.

1.1 Background and Motivation

Fairness learning has received massive attention in recent years. It turns out learning a fair classifier, in most cases, is equivalent to auditing [20, 23, 18]. In particular, if auditing is possible, learning a fair classifier is easy. There are many successful examples of fairness learning with auditing over a relatively small number of predetermined subgroups [1, 29]. However, a small number of predetermined subgroups, in many cases, is not enough to cover all the natural subgroups.

▶ **Example 1.** In the court case "DeGraffenreid v General Motors" [6], five Black women brought suit against General Motors for its discrimination against the group of Black women. Although no sex discrimination was revealed, the evidence showed that Black women hired after 1970 were discriminated against by the company's seniority system. Such discrimination can be better illustrated by an example shown in Table 1. In particular, the hiring rate of a company could seemingly be fair in terms of gender or race alone, but clearly discriminates against the subgroups of white men and black women. The court rejected the plaintiffs' attempt to bring a suit not on behalf of Blacks or women, but specifically on behalf of Black women. In the ruling, in favor of the defendant, the judge was specifically concerned about the proliferation of protected classes.

Table 1 an example of discrimination against subgroups.

	men	women	total
black	50	0	50
white	0	50	50
total	50	50	100

More generally, a classifier may appear to be fair on each individual attribute, e.g., gender, race, age, incomes, etc., and yet perform unfairly on subgroups defined on multiple attributes, i.e., the conjunction of such attributes. In the case of *DeGraffenreid v General Motors*, it is the conjunction of race and gender being discriminated against. The possible number of the conjunctions grows exponentially as the number of the "protected" attributes increases.

Thereafter, [20] proposed more general notions of statistical fairness that require auditing over subgroups defined on simple combinations of data features. Specifically, such combinations of features can be any simple representations, such as conjunctions and halfspaces, which, however, can generate exponentially many subgroups. They also showed that the problem of auditing subgroups defined by such simple representation is as hard as "weak agnostic learning" in the standard "distribution-free" setting [17, 22]. While the problem of distribution-free weak agnostic learning is widely believed to be computationally intractable [22, 12], its hardness does not necessarily hold for specific distribution families. Thus, it is natural to consider auditing using distribution-specific agnostic learning approaches as agnostic learning is a much more extensively studied problem. However, it turns out there are still obstacles remaining for doing so.

1.2 Challenges of Auditing through Agnostic Learning

The main challenge that prevents us from applying existing agnostic learning techniques to perform auditing based on the reduction by [20] is that it is formulated in terms of weak agnostic learning, that is, finding classifiers with error rates that are nonnegligibly better than guessing, and correspondingly weak auditing guarantees. In particular, the approximation guarantees we obtain for distribution-specific agnostic learning yield vacuous guarantees for weak learning. When we have guarantees for arbitrary distributions, "boosting" [28] enables us to obtain high accuracy from such weak learners. Unfortunately, these techniques require re-weighting the data examples after which the distribution-specific properties may no longer hold.

One might hope to dodge this issue by casting the problem of finding a harmed subgroup as a Mixed-Integer Program and using solvers that, though they lack polynomial-time guarantees, obtain adequate performance in practice. In such an approach, the failure of the solver to find a feasible solution to the optimization problem is taken as the proof that the classifier is fair. Unfortunately, these solvers owe their speed in part to a lack of soundness, both due to numerical issues [5] and the complexity of the heuristics used to prune the search [2, 14], and it remains a current research challenge to obtain acceptable performance (using the various advanced techniques employed by commercial solvers) while retaining the guarantee that the solver correctly reports infeasibility [4]. In any case, the works by [20, 21] and [24] that empirically studied these approaches to obtaining fair classifiers used linear regression as a proxy for the agnostic learning or cost-sensitive classification subroutines. Unfortunately, these heuristics do not even provide in-principle guarantees.

In this paper, we will show auditing general halfspace subgroups is hard even for data with a Gaussian distribution, and present an alternative auditing approach for subgroups determined by homogeneous halfspaces with provable guarantees.

1.3 Our Contribution

Our first contribution is a more careful analysis of the relationship between auditing and agnostic learning: Given a fixed positive classification rate, the harm (w.r.t. statistical parity) suffered by a subgroup is affinely related to the error rate of the subgroup indicator. Thus, a solution to the agnostic learning problem directly gives a harmed subgroup. Note that whereas the fairness objective refers to conditioning on a group, which generally doesn't preserve a distributional assumption, agnostic learning instead refers to the accuracy under that "nice" distribution, and hence is easier to analyze. Also note that under a standard normal distribution, the subclass of halfspaces with a fixed positive classification rate is given by the halfspaces with unit normal vectors and the same threshold.

▶ Remark 2. Our reduction to learning halfspaces with fixed positive classification rates can achieve arbitrarily high precision auditing and does not rely on re-weighting data examples or make any assumptions on the potentially unfair classifiers. This enables the use of the existing distribution-specific agnostic learning methods for auditing.

Based on the reduction and a inspiration from Diakonikolas et al. [7], our second major contribution is a lower bound on the unfairness detectable when auditing for halfspace subgroups under Gaussian distributions by reducing the problem of continuous Learning With Errors (cLWE) to auditing. Our hardness results include both multiplicative and additive forms. More interestingly, we can further show that even "nonconstructive auditing" is hard, where we do not need to exhibit a discriminated subgroup for a failed audit.

For our algorithmic results, we will present a general auditing framework given an oracle for (distribution-specific) agnostic learning. Also, we give a randomized PTAS auditing algorithm for subgroups determined by homogeneous halfspaces under Gaussian data by applying the method from Diakonikolas et al. [8].

▶ Remark 3. We stress that a PTAS for auditing subgroups defined by homogeneous halfspaces for Gaussian distributions is, in fact, the best guarantee we know so far, hence, not trivial.

At first blush, the reliance on a (prima facie unverifiable) distributional assumption for the analysis of our auditing algorithm may seem to be at odds with our desire to certify the fairness of a classifier. Nevertheless, a line of recent works by Rubinfeld and Vasilyan [27] and Gollakota et al. [15] have shown that the properties of the data that are crucial to these algorithms for distribution-specific learning of halfspaces *can be verified*. Thus, these methods give a way of certifying fairness for families of nice distributions: so long as the data passes these tests and the audit reveals no subgroup that is significantly harmed, we may *guarantee that the classifier is fair*.

This paper will be organized as follows. Some necessary background for our arguments are given in Section 2. We will present the main reduction from auditing to agnostic learning in Section 3. Then, we will show the hardness results in Section 4. Section 5 will present our auditing framework as well as the distribution-specific PTAS algorithm. Finally, we will discuss the limitations of our approach and suggest directions for future work.

1.4 Related Work

Many authors have considered the problem of ensuring fairness in classification, and Barocas et al. [3] give a good overview of the broader area. In particular, there are alternatives to the statistical, group-fairness notions we are considering, for example individual-level fairness as proposed by Dwork et al. [11], or based on causal modeling, such as the "counterfactual" fairness notion proposed by Kusner et al. [25]. We cannot do justice to the breadth of literature and philosophical issues here, and we strongly encourage the interested reader to consult Barocas et al. The group-fairness notions we consider have their roots in the game-theory-based approach of Kearns et al. [20] for learning representations with subgroup fairness by assuming there exists an efficient oracle for auditing. A follow-up study [21] evaluated their algorithm on real-world datasets. Hébert-Johnson et al. [18] showed a method of obtaining "multi-accurate" representations by assuming the existence of an efficient auditing oracle. Further, Kim et al. [23] proposed a variant of statistical fairness called "multi-fairness," which allows them to efficiently learn a multi-fair classifier with querying "relative fairness" of data pairs. As we discussed previously, the auditing oracles in these works were provided by using linear regression as a heuristic for the optimal halfspace, which does not provide guarantees. They also did not consider auditing for specific families of distributions. On the other hand, the works on agnostic learning for specific families of distributions, e.g., [19, 9, 8, 10, 13] do not consider how their techniques may be applied to the subgroup fairness auditing problem.

2 Preliminaries

We use lowercase bold font characters to represent real vectors and subscripts to index the coordinates of each vector, e.g., $\mathbf{x}_i$ represents the i-th coordinate of vector $\mathbf{x}$. We denote the l_p-norm by $\|\mathbf{x}\|_p = \left(\sum_i \mathbf{x}_i^p\right)^{1/p}$, and $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$. We model each individual as a vector of protected attributes, i.e., $\mathbf{x} \in \mathcal{X}$.

Further, the probability of an event under a distribution $\mathcal{D}$ is denoted by $\Pr_{\mathbf{x}\sim\mathcal{D}}\{\cdot\}$. $\mathcal{N}(0,\mathrm{I})$ denotes a standard normal distribution, where I represents the identity matrix. For simplicity of notation, we may use $\mathcal{N},\mathcal{N}_\sigma$ instead of $\mathcal{N}(0,\mathrm{I}),\mathcal{N}(0,\sigma^2\mathrm{I})$ or even drop $\mathcal{D}$ and $\mathcal{N}$ from the subscript when it is clear from the context.

▶ **Fact 4** (Rotational Invariance). *For any real vector* $\mathbf{u}$*, if* $\mathbf{x}\sim\mathcal{N}(0,\mathrm{I})$*, then* $\bar{\mathbf{u}}^\top\mathbf{x}\sim\mathcal{N}(0,1)$.

To understand the problem of fairness auditing, it is necessary to define fairness or unfairness precisely. In this work, we focus on the notion of Statistical Parity Subgroup Fairness (SPSF). Formally, we have the following definition.

▶ **Definition 5** (Statistical Parity Subgroup Fairness). *Fix any binary classifier* $c\in\mathcal{C}$ *such that* $c:\mathbb{R}^d\to\{-1,+1\}$*, data distribution* $\mathcal{D}$*, collection of subgroups* $\mathcal{G}$*, and parameter* $\gamma\in[0,1]$*. Define*

$$d_{\mathcal{D}}(c,g)=\Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x})=1\}-\Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x})=1\mid\mathbf{x}\in g\} \tag{1}$$

We say that c does not satisfy γ-statistical parity fairness (or is γ-unfair) with respect to $\mathcal{D}$ and $\mathcal{G}$*, if* $\exists g\in\mathcal{G}$ *such that*

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{\mathbf{x}\in g\}\,|d_{\mathcal{D}}(c,g)|\geq\gamma \tag{2}$$

Equation (1) is a straightforward way to quantify how much the positive classification rate within a subgroup deviates from that of the overall population. The weighting by the size of the group (i.e., $\Pr_{\mathbf{x}\sim\mathcal{D}}\{\mathbf{x}\in g\}$) is a concession to address the statistical issues that arise with estimating d on small groups: we cannot escape that our empirical estimates are less accurate as the size shrinks. Our approach makes no assumptions on the form of the function c; note therefore, that by replacing c with other functions of $\mathbf{x}$, such as whether a given classifier agrees with a given label, or whether the classifier makes a false-positive error, our results will immediately extend to other standard notions of statistical subgroup fairness. The goal of fairness auditing is to develop an "auditing algorithm" to efficiently find such a certificate $g\in\mathcal{G}$ for any $c\in\mathcal{C}$ with sample access to $\mathcal{D}$, formalized as follows.

▶ **Definition 6** (Constructive Auditing [20]). *Fix a collection of group indicators* $\mathcal{G}$ *over the protected features, and any* $\delta,\gamma,\gamma'\in(0,1)$ *such that* $\gamma'\leq\gamma$*. A* constructive (γ,γ')-auditing algorithm for $\mathcal{G}$ with respect to distribution $\mathcal{D}$ *is an algorithm* $\mathcal{A}$ *such that for any classifier* h*, when given access the joint distribution* $(\mathcal{D},h(\mathcal{D}))$*,* $\mathcal{A}$ *runs in time* $\mathrm{poly}(1/\gamma',\log(1/\delta))$*, and with probability* $1-\delta$*, outputs a* γ'*-unfair certificate for* h *whenever* h *is* γ*-unfair with respect to* $\mathcal{D}$ *and* $\mathcal{G}$*. If* h *is* γ'*-fair,* $\mathcal{A}$ *will output "fair".*

Moreover, we will consider a more general type of auditing task, called "non-constructive auditing", where the algorithms are only required to tell if a discriminated subgroup exists.

▶ **Definition 7** (Non-constructive Auditing). *Under the same setting as Definition 6, a* non-constructive (γ,γ')-auditing algorithm for $\mathcal{G}$ with respect to distribution $\mathcal{D}$ *is an algorithm* $\mathcal{A}$ *such that for any classifier* h*, when given access the joint distribution* $(\mathcal{D},h(\mathcal{D}))$*,* $\mathcal{A}$ *runs in time* $\mathrm{poly}(1/\gamma',\log(1/\delta))$*, and with probability* $1-\delta$*, claims* h *is* γ'*-unfair whenever* h *is* γ*-unfair with respect to* $\mathcal{D}$ *and* $\mathcal{G}$*. If* h *is* γ'*-fair,* $\mathcal{A}$ *will output "fair".*

In this work, we will mainly focus on subgroups defined on halfspaces, a.k.a. linear threshold functions (LTF) over a d-dimensional real domain. Formally:

▶ **Definition 8** (Halfspaces). *The class of* halfspaces *over* $\mathbb{R}^d$ *is defined as* $\mathcal{H}^d := \{\mathbf{x} \mapsto \text{sgn}(\mathbf{v}^\top \mathbf{x} - t) \mid \mathbf{x}, \mathbf{v} \in \mathbb{R}^d, t \in \mathbb{R}\}$ *where* $\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & \text{otherwise} \end{cases}$. *In particular, the class of* homogeneous halfspaces *can be defined as* $\{\mathbf{x} \mapsto \text{sgn}(\mathbf{v}^\top \mathbf{x}) \mid \mathbf{x}, \mathbf{v} \in \mathbb{R}^d\}$.

Since our reduction involves the subclass of halfspace subgroups of a fixed size, we give the formal definition of it as follows.

▶ **Definition 9** (Fixed-size Halfspaces). *We use* $\mathcal{H}^d$ *to represent the collection of all halfspaces in* $\mathbb{R}^d$. *Then, for any arbitrary distribution* $\mathcal{D}$ *over* $\mathbb{R}^d$, *we define the collection of all* halfspaces with the same (relative) density μ *as*

$$\mathcal{H}^{\mathcal{D}}_{\mu} := \{h \in \mathcal{H}^d \mid \Pr_{\mathbf{x} \in \mathcal{D}}\{h(\mathbf{x}) = 1\} = \mu\} \tag{3}$$

In particular, the class of homogeneous halfspaces for a mean-0 Gaussian distribution is $\mathcal{H}^{\mathcal{N}(0,\Sigma)}_{1/2}$.

For conciseness, we may abbreviate $\Pr\{f(\mathbf{x}) = 1\}$ and $\Pr\{f(\mathbf{x}) = -1\}$ to simply $\Pr\{f\}$ and $\Pr\{\neg f\}$ for any binary output functions $f : \mathcal{X} \to \{-1, +1\}$ in the rest of the paper.

To state the hardness results, we denote $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$, $\mathbb{Z}_q := \{0, 1, \ldots, q-1\}$, $\mathbb{R}_q := [0, q)$, and $\text{mod}_q : \mathbb{R}^d \to \mathbb{R}_q$ for the unique translation of the input by $q\mathbb{Z}^d$ to $\mathbb{R}_q$ for $q \in \mathbb{N}$. The hardness of distribution-specific auditing is based on the assumption that the problem of "Learning With Errors" (LWE) is computationally intractable. Informally speaking, in the problem of LWE, we are given labelled examples from two hypothesis cases. In one case, the labels are biased by some secret vector, while, in another case, the labels are generated uniformly at random. We wish to distinguish between these cases. We formally define the problem of LWE [26], following [7]:

▶ **Definition 10** (Learning With Errors). *For* $m, d \in \mathbb{N}$, $q \in \mathbb{R}_+$, *let* $\mathcal{D}_{sample}, \mathcal{D}_{secret}, \mathcal{D}_{noise}$ *be distributions on* $\mathbb{R}^d, \mathbb{R}^d, \mathbb{R}$ *respectively. In the LWE*$(m, \mathcal{D}_{sample}, \mathcal{D}_{secret}, \mathcal{D}_{noise}, \text{mod}_q)$ *problem, with* m *independent samples* $(\mathbf{x}, y)$, *we want to distinguish between the following two cases:*

- ***Alternative hypothesis****:* $(\mathbf{x}, y)$ *is generated as* $y = \text{mod}_q(\mathbf{s}^\top \mathbf{x} + z)$, *where* $\mathbf{x} \sim \mathcal{D}_{sample}, \mathbf{s} \sim \mathcal{D}_{secret}, z \sim \mathcal{D}_{noise}$.
- ***Null hypothesis****:* y *is sampled uniformly at random on the support of its marginal distribution in the alternative hypothesis, independent of* $\mathbf{x} \sim \mathcal{D}_{sample}$.

An algorithm is said to be able to solve the LWE problem with Δ advantage *if the probability that the algorithm outputs "alternative hypothesis" is* Δ *larger than the probability that it outputs "null hypothesis" when the given data is sampled from the alternative hypothesis distribution.*

This problem is widely believed to be computationally hard, formalized as follows.

▶ **Assumption 11** (Sub-exponential LWE Assumption). *For* $q, \kappa \in \mathbb{N}, \alpha \in (0, 1)$ *and* $C > 0$ *being a sufficiently large constant, the problem LWE*$(2^{O(n^\alpha)}, \mathbb{Z}_q^d, \mathbb{Z}_q^d, \mathcal{N}_\sigma, \text{mod}_q)$ *with* $q \leq d^\kappa$ *and* $\sigma = C\sqrt{d}$ *cannot be solved in* $2^{O(d^\alpha)}$ *time with* $2^{O(-d^\alpha)}$ *advantage.*

3 From Auditing To Agnostic Learning

In this section, we describe our reduction from auditing to agnostic learning. In addition, we give a lower bound for fairness auditing under Gaussian distributions.

We are considering the auditing problem w.r.t. SPSF as in Definition 5, which naturally rules out the statistically small subgroups. Indeed, if the probability of accessing the data of certain sub-population is exponentially small, it is statistically hard to even estimate their deviation. Therefore, it makes sense to just consider the collection of subgroups $\mathcal{G}$ that are statistically large enough, e.g., $\Pr\{\mathbf{x} \in g\} = \Theta(1)$ for $\mathbf{x} \in \mathbb{R}^d$.

Based on the observation, the following optimization program, $\mathcal{P}^{\mathcal{D}}_{a,b}(\mathcal{G})$, can capture the most unfair subgroup which is also statistically significant enough. That is

$$\begin{aligned} \max_{g \in \mathcal{G}} \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{\mathbf{x} \in g\} \, |d_{\mathcal{D}}(c, g)| \\ s.t. \quad & a \leq \Pr_{\mathbf{x} \in \mathcal{D}} \{\mathbf{x} \in g\} \leq b \end{aligned} \tag{4}$$

for some constants $0 < a \leq b < 1$.

Furthermore, if we only consider the subgroups represented by halfspaces, i.e., $\mathcal{G} \equiv \mathcal{H}^d$, there exists a simple reduction from $\mathcal{P}^{\mathcal{D}}_{a,b}(\mathcal{H}^d)$ to agnostic learning that, in particular, preserves the properties of the data distribution. We show our reduction as the following theorem.

▶ **Theorem 12** (Main Reduction). *Given any binary classifier $c : \mathbb{R}^d \to \{-1, +1\}$, and a data distribution $\mathcal{D}$ over $\mathbb{R}^d$ whose 1-dimensional marginals have continuous cumulative distribution functions, if there exists an efficient algorithm for learning $\mathcal{H}^{\mathcal{D}}_{\mu}$ in the agnostic model on distribution $\mathcal{D}$, then there is an efficient auditing algorithm for c on subgroups represented by $\mathcal{H}^d$ over distribution $\mathcal{D}$.*

We delay the proof of the above theorem to the end of this section, and show two fundamental hurdles we need to overcome in order to prove Theorem 12.

▶ Remark 13. While learning from a representation class like $\mathcal{H}^{\mathcal{D}}_{\mu}$ may seems to be hard at a first glance, there are actually examples [10] of learning $\mathcal{H}^{\mathcal{D}}_{\mu}$ in an agnostic setting under Gaussian data.

Instead of starting from the optimization problem (4), it turns out that solving a sequence of simpler optimization problems suffices to certify the γ-unfairness as stated in Definition 5. We state the equivalence in the following proposition. Its proof is deferred to the appendix.

▶ **Proposition 14.** *Consider any binary classifier $c : \mathbb{R}^d \to \{-1, +1\}$, any data distribution $\mathcal{D}$ over $\mathbb{R}^d$ whose 1-dimensional marginals have continuous cumulative distribution functions, and any $0 < a \leq b < 1$. For each pair of non-negative integers $k < n$, let $\mathcal{P}^{\mathcal{D}}_{a,b}(k, n)$ denote the optimization program*

$$\begin{aligned} \max_{h \in \mathcal{H}^d} \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} \, |d_{\mathcal{D}}(c, h)| \\ s.t. \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} = a + \frac{k(b-a)}{n}. \end{aligned}$$

Let h^* be a global optimizer of $\mathcal{P}^{\mathcal{D}}_{a,b}(\mathcal{H}^d)$, as defined in (4), and let $\gamma^* = \Pr\{h^*\} \, |d_{\mathcal{D}}(c, h^*)|$. For each $k = 0, \ldots, n$, let h^*_k be a global optimizer of $\mathcal{P}^{\mathcal{D}}_{a,b}(k, n)$. Then

$$\max_k \Pr\{h^*_k\} \, |d_{\mathcal{D}}(c, h^*_k)| \geq \gamma^* - \frac{2(b-a)}{n}.$$

The reason why this proposition is so crucial is that it allows us to solve a simpler optimization problem without compromising the guarantee. Being able to fix $\Pr\{h(\mathbf{x}) = 1\}$ as a constant will significantly simplify the overall optimization as it reduces the degree of

the optimization objective. In fact, it is because we can optimize $\Pr\{h(\mathbf{x}) = 1\}\,|d_{\mathcal{D}}(c,h)|$ over $\mathcal{H}_\mu^{\mathcal{D}}$ instead of $\mathcal{H}^d$ that we can conduct the reduction from auditing to agnostic learning.

The following lemma shows a direct relationship between the unfairness level and the classification error.

▶ **Lemma 15.** *Given any binary classifier $c : \mathcal{X} \to \{-1, +1\}$, a data distribution $\mathcal{D}$ over $\mathcal{X}$ and a collection of subgroups $g \in \mathcal{G}$ such that $g : \mathcal{X} \to \{-1, +1\}$, we have*

$$2\Pr\{g\}d_{\mathcal{D}}(c,g) = \Pr\{\neg c\}\Pr\{\neg g\} + \Pr\{c\}\Pr\{g\} - \Pr\{c(\mathbf{x}) = g(\mathbf{x})\}$$

for $\mathbf{x} \sim \mathcal{D}$.

Proof. By the law of total probability, we have

$$\Pr\{c \cap g\} = \Pr\{g\} - (\Pr\{\neg c\} - \Pr\{\neg c \cap \neg g\}).$$

which along with Definition 5 gives

$$\begin{aligned} d_{\mathcal{D}}(c,g) =& \Pr\{c\} - \Pr\{c \mid g\} \\ =& \frac{\Pr\{c\}\Pr\{g\} - \Pr\{c \cap g\}}{\Pr\{g\}} \\ =& \frac{\Pr\{\neg c\}\Pr\{\neg g\} - \Pr\{\neg c \cap \neg g\}}{\Pr\{g\}}. \end{aligned} \tag{5}$$

Summing up the two different forms of $d_{\mathcal{D}}(c,g)$ results to

$$\begin{aligned} 2d_{\mathcal{D}}(c,g) =& \frac{\Pr\{\neg c\}\Pr\{\neg g\} - \Pr\{\neg c \cap \neg g\}}{\Pr\{g\}} + \frac{\Pr\{c\}\Pr\{g\} - \Pr\{c \cap g\}}{\Pr\{g\}} \\ =& \frac{\Pr\{\neg c\}\Pr\{\neg g\} + \Pr\{c\}\Pr\{g\} - (\Pr\{\neg c \cap \neg g\} + \Pr\{c \cap g\})}{\Pr\{g\}} \end{aligned} \tag{6}$$

Notice that, because $c \cap g$ and $\neg c \cap \neg g$ are two disjoint events, we have

$$\begin{aligned} \Pr\{c(\mathbf{x}) = g(\mathbf{x})\} =& \Pr\{(c \cap g) \cup (\neg c \cap \neg g)\} \\ =& \Pr\{c \cap g\} + \Pr\{\neg c \cap \neg g\} \end{aligned}$$

Plugging it back in to Equation (6) produces the desired result. ◀

This immediately implies a duality between SPSF auditing and agnostic learning as follows.

▶ **Corollary 16.** *Given any binary classifier $c : \mathbb{R}^d \to \{-1, +1\}$, a data distribution $\mathcal{D}$ and a collection of halfspaces $\mathcal{H}_\mu^{\mathcal{D}}$ over $\mathbb{R}^d$, we have the following two properties*
(1) *$d_{\mathcal{D}}(c,h^*) \geq d_{\mathcal{D}}(c,h), \forall h \in \mathcal{H}_\mu^{\mathcal{D}}$ if and only if $h^* = \operatorname{argmin}_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x}) = h(\mathbf{x})\}$*
(2) *$d_{\mathcal{D}}(c,h^*) \leq d_{\mathcal{D}}(c,h), \forall h \in \mathcal{H}_\mu^{\mathcal{D}}$ if and only if $h^* = \operatorname{argmax}_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x}) = h(\mathbf{x})\}$*

Proof. Because $\Pr\{c\}$ is a constant and $\Pr\{h\} = \mu, \forall h \in \mathcal{H}_\mu^{\mathcal{D}}$ by Definition 9, $d_{\mathcal{D}}(c,h)$ is simply an affine transformation of $\Pr\{c(\mathbf{x}) = h(\mathbf{x})\}$ for a fixed μ by Lemma 15, which implies the desired results. ◀

Proposition 14 tells us that solving $\mathcal{P}_{a,b}^{\mathcal{D}}(k,n)$ for $k = 0, \ldots, n$ would give us a good enough approximation to the maximum unfairness level, of course, with a large enough n. Therefore, we just need to further show that solving each $\mathcal{P}_{a,b}^{\mathcal{D}}(k,n)$ is equivalent to learning $\mathcal{H}_\mu^{\mathcal{D}}$ to complete the reduction.

Formally, because $\mathcal{P}^{\mathcal{D}}_{a,b}(k,n)$ can be equivalently written as

$$\max_{h\in\mathcal{H}^{\mathcal{D}}_{\mu}} \Pr_{\mathbf{x}\in\mathcal{D}}\{h(\mathbf{x})=1\}\,|d_{\mathcal{D}}(c,h)| \tag{7}$$

for some $\mu = a + k(b-a)/n$, it suffices to prove the following theorem.

▶ **Lemma 17.** *Given any binary classifier $c:\mathbb{R}^d\to\{-1,+1\}$, a data distribution $\mathcal{D}$ and a collection of halfspaces $\mathcal{H}^{\mathcal{D}}_{\mu}$ over $\mathbb{R}^d$ such that*

$$\mathrm{opt}_{\min} \le \Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x})=h(\mathbf{x})\} \le \mathrm{opt}_{\max}$$

for all $h\in\mathcal{H}^{\mathcal{D}}_{\mu}$, if $h_{\mathbf{v}}, h_{\mathbf{u}}\in\mathcal{H}^{\mathcal{D}}_{\mu}$ satisfy that $\Pr\{c(\mathbf{x})=h_{\mathbf{v}}(\mathbf{x})\}\le\mathrm{opt}_{\min}+2\epsilon$ as well as $\Pr\{c(\mathbf{x})=h_{\mathbf{u}}(\mathbf{x})\}\ge\mathrm{opt}_{\max}-2\epsilon$, we have either

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{h_{\mathbf{v}}(\mathbf{x})=1\}\left|d_{\mathcal{D}}(c,h_{\mathbf{v}})\right| \ge \gamma^*-\epsilon \tag{8}$$

or

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{h_{\mathbf{u}}(\mathbf{x})=1\}\left|d_{\mathcal{D}}(c,h_{\mathbf{u}})\right| \ge \gamma^*-\epsilon \tag{9}$$

where $\gamma^* = \max_{h\in\mathcal{H}^{\mathcal{D}}_{\mu}}\Pr_{\mathbf{x}\sim\mathcal{D}}\{h(\mathbf{x})=1\}\,|d_{\mathcal{D}}(c,h)|$.

Proof. By the proof of Lemma 15, we have

$$2\Pr\{h\}\,|d_{\mathcal{D}}(c,h)| = |\underbrace{\Pr\{\neg c\}\Pr\{\neg h\}-\Pr\{\neg c\cap\neg h\}}_{I_1}+\underbrace{\Pr\{c\}\Pr\{h\}-\Pr\{c\cap h\}}_{I_2}|$$

Let $h^*\in\mathcal{H}^{\mathcal{D}}_{\mu}$ be such that $\Pr\{h^*\}\,|d_{\mathcal{D}}(c,h^*)|=\gamma^*$. Then for I_2, we have

$$\begin{aligned} I_2 =& (\Pr\{c\}-\Pr\{c\mid h^*\}+\Pr\{c\mid h^*\})\Pr\{h\}-\Pr\{c\cap h\}\\ =& \Pr\{h^*\}d_{\mathcal{D}}(c,h^*)+\Pr\{c\cap h^*\}-\Pr\{c\cap h\}\end{aligned}$$

where the last equation is because $h^*\in\mathcal{H}^{\mathcal{D}}_{\mu}$, then $\Pr\{h\}=\Pr\{h^*\}=\mu$ by Definition 9.

Similarly, for I_1, we can write

$$\begin{aligned} I_1 =& \Pr\{\neg h^*\}(\Pr\{\neg c\}-\Pr\{\neg c\mid\neg h^*\})+\Pr\{\neg c\cap\neg h^*\}-\Pr\{\neg c\cap\neg h\}\\ =& \Pr\{h^*\}d_{\mathcal{D}}(c,h^*)+\Pr\{\neg c\cap\neg h^*\}-\Pr\{\neg c\cap\neg h\}\end{aligned}$$

where the last equation follows because we have shown in the proof of Lemma 15 that $d_{\mathcal{D}}(c,h^*)=\Pr\{\neg h^*\}(\Pr\{\neg c\}-\Pr\{\neg c\mid\neg h^*\})/\Pr\{h^*\}$.

Combining I_1 and I_2 will result to

$$\begin{aligned}\Pr\{h\}\,|d_{\mathcal{D}}(c,h)| =& |\Pr\{h^*\}d_{\mathcal{D}}(c,h^*)+\frac{\Pr\{c(\mathbf{x})=h^*(\mathbf{x})\}-\Pr\{c(\mathbf{x})=h(\mathbf{x})\}}{2}|\\ \ge& \gamma^*-\frac{|\Pr\{c(\mathbf{x})=h^*(\mathbf{x})\}-\Pr\{c(\mathbf{x})=h(\mathbf{x})\}|}{2}\end{aligned}$$

by triangle inequality. Further, since h^* maximizes $|d_{\mathcal{D}}(c,h)|$, it either maximizes or minimizes $d_{\mathcal{D}}(c,h)$. Then, by Corollary 16, we know

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{c(\mathbf{x})=h^*(\mathbf{x})\}\in\{\mathrm{opt}_{\min},\mathrm{opt}_{\max}\}$$

which implies either

$$\left|\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h_{\mathbf{v}}(\mathbf{x})\}\right| \leq 2\epsilon$$

or

$$\left|\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h_{\mathbf{u}}(\mathbf{x})\}\right| \leq 2\epsilon$$

Therefore, the proof is completed. ◀

▶ Remark 18. We emphasize that it is necessary for us to consider the guarantee of agnostic learning in a additive form rather than multiplicative form. Although Corollary 16 shows that the classification error, $\Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$, and the unfairness level, $\Pr\{h\}\,|d_{\mathcal{D}}(c,h)|$, are dual to each other over $\mathcal{H}_{\mu}^{\mathcal{D}}$, the affine relationship between them prohibits obtaining a guarantee on the unfairness from a multiplicative error. This also explains why the guarantee provided by [10] does not fit in our analysis.

Now we are ready to prove Theorem 12.

Proof of Theorem 12. To solve the auditing problem, we just need to solve the sequence of optimization problems, $\{\mathcal{P}_{a,b}^{\mathcal{D}}(k,n) \mid k = 0,\ldots,n\}$ as described in Proposition 14. We can solve each $\mathcal{P}_{a,b}^{\mathcal{D}}(k,n)$ with an additive error ϵ by calling the given oracle of learning halfspaces with the same strategy specified in Lemma 17. Eventually, we solve all of these optimization problems with an $2(b-a)/n + \epsilon$ additive error and a running time of $O(n)$ factor overhead compared with that of the oracle. ◀

4 Intractability Of Auditing Under Gaussian Data

In this section, we will show that the problem of auditing halfspaces subgroups under a Gaussian distribution is computationally hard in two forms: the multiplicative form and additive form. To do so, we first show that distinguishing between fair and unfair cases with respect to halfspace subgroups for Gaussian data is hard. Then, the hardness of auditing will follow as corollaries.

4.1 Indistinguishability Of Unfairness

We claim it is computationally hard to distinguish between halfspace subgroups that are evenly fair and halfspace subgroups among which there exists a slightly unfair subgroup with significant advantage.

▶ **Theorem 19.** *Under Assumption 11, for any $d \in \mathbb{N}$, any constants $\alpha \in (0,1), \beta \in \mathbb{R}_+$, and any $\log^{\beta} d \leq k \leq cd$ where c is a sufficiently small constant, there is no algorithm that runs in time $d^{O(k^{\alpha})}$ and distinguishes between the following two cases of a joint distribution $\mathcal{D}$ of $(\mathbf{x}, c(\mathbf{x}))$ supported on $\mathbb{R}^d \times \{-1,+1\}$ with marginal $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(0,\mathrm{I})$, with $d^{-O(k^{\alpha})}$ advantage:*

- **(i)** ***Alternative Hypothesis**: There exist non-negligibly unfair halfspace subgroups, specifically $\exists h \in \mathcal{H}^d, \Pr_{\mathcal{D}}\{h(\mathbf{x}) = 1\}|d_{\mathcal{D}}(c,h)| = \Omega(1/\sqrt{k \log d})$.*
- **(ii)** ***Null Hypothesis**: All halfspace subgroups are perfectly fair, i.e., $\Pr_{\mathcal{D}}\{h(\mathbf{x}) = 1\}|d_{\mathcal{D}}(c,h)| = 0, \forall h \in \mathcal{H}^d$.*

The above theorem simply states that the closer the unfairness level of the alternative hypothesis is to zero ($k \log d$ is large), the harder it is to distinguish between these two cases, where the hardness is reflected on the running time $d^{O(k^{\alpha})}$. Hence, if we restrict the running

time to a certain order, there is a limitation on how large $k \log d$ can be for someone to be able to distinguish between them with a significant enough advantage. It is this observation that allows us to prove the hardness of auditing in the next section.

The idea behind the proof of this theorem is to observe that the data generated in the two hypotheses in certain LWE instances can be reduced to binary labelled ones through rounding. With such a reduction, the distribution from the null hypothesis case of LWE will produce perfectly fair data, while the distribution from alternative hypothesis will yield slightly biased labels where a unfair halfspace subgroup therefore exists. Thus, if we can distinguish between the fair case from the unfair case with some marginal error, we can solve the LWE problem. We defer the formal proof to the appendix.

4.2 Auditing With Small Error Is Hard

We now show that the hardness of distinguishability implies the hardness of auditing with both multiplicative error and additive error.

Suppose an auditing algorithm is guaranteed to return us a γ'-unfair certificate (a halfspace) given a γ-unfair classifier c, where $\gamma' \leq \gamma \leq 1$. The following corollaries show that γ' can never be close to γ.

▶ **Corollary 20** (multiplicative form). *Given Assumption 11, there is no polynomial-time* $1/\mathrm{poly}(d)$*-approximation algorithm for constructive auditing for halfspace subgroups under Gaussian marginals in* $\mathbb{R}^d$*.*

Proof. Suppose there exists an auditing algorithm that guarantees to return a $\delta\gamma$-unfair certificate given a γ-unfair collection of halfspace subgroup and access to data with a Gaussian marginal, where $\delta \in (0, 1)$.

For the alternative hypothesis case as described in Theorem 19, given a $1/\sqrt{k \log d}$-unfair collection of halfspace subgroups, we run such an algorithm to obtain a $\delta/\sqrt{k \log d}$-unfair certificate, i.e., a halfspace h such that $\Pr_{\mathbf{x}\sim\mathcal{N}}\{h(\mathbf{x}) = 1\}|d_{\mathcal{N}}(c, h)| \geq \delta/\sqrt{k \log d}$. By the Hoeffding Bound, we can verify that the empirical estimation of $\Pr_{\mathbf{x}\sim\mathcal{N}}\{h(\mathbf{x}) = 1\}|d_{\mathcal{N}}(c, h)|$ is ε_1-close to $\delta/\sqrt{k \log d}$ with high probability by drawing $O(1/\varepsilon_1^2)$ examples from the distribution constructed in the alternative hypothesis case.

For the null hypothesis case, with the same argument, we can verify there is no ε_2-unfair subgroup with high probability given $O(1/\varepsilon_2^2)$ examples from the distribution in the null hypothesis case.

Suppose $\delta = \Omega(1/\mathrm{poly}(d))$, notice that we only need $\varepsilon_1, \varepsilon_2$ to be $O(1/\mathrm{poly}(d))$ to ensure $\delta/\sqrt{k \log d} - \varepsilon_1 > \varepsilon_2$. However, this implies that our auditing algorithm can distinguish between the two cases in Theorem 19 with high probability and only runs in polynomial time, which contradicts to the hardness assumption. ◀

▶ **Corollary 21** (additive form). *Given Assumption 11, for any constants* $\alpha \in (0, 1), \beta \in \mathbb{R}_+$*, and any* $C/\sqrt{d \log d} \leq \epsilon \leq c'/\log^{(1+\beta)/2} d$ *where* C *is a sufficiently large constant and* c' *is a sufficiently small constant, no auditing algorithm can return a unfair certificate for halfspace subgroups in* $\mathbb{R}^d$ *with an additive error* ϵ *under Gaussian marginals and runs in time* $d^{O(1/(\epsilon^2 \log d)^\alpha)}$*.*

Proof. Suppose there exists an auditing algorithm that guarantees to return a $\gamma - \epsilon$-unfair certificate given a γ-unfair collection of halfspace subgroups and access to data with a Gaussian marginal, where $\epsilon \in (0, 1)$.

Similar to the proof of Corollary 20, given a $1/\sqrt{k \log d}$-unfair collection of halfspace subgroups, we run such an algorithm to obtain a $(1/\sqrt{k \log d} - \epsilon)$-unfair certificate. Observe that, if $\epsilon = c'/\sqrt{k \log d}$ for some sufficiently small constant c', we can solve the testing problem

in Theorem 19 within time $d^{O(k^\alpha)}$ by running this algorithm as well as drawing enough examples to estimate the unfairness of the returned certificates from the two cases respectively. On the other hand, given $\epsilon = c'/\sqrt{k \log d}$, we can rewrite $d^{O(k^\alpha)} = d^{O(1/(\epsilon^2 \log d)^\alpha)}$.

However, Theorem 19 tells that the above case is impossible for any $C/\sqrt{d \log d} \le \epsilon \le c'/\log^{(1+\beta)/2} d$, where C is a sufficiently large constant. ◀

Besides the general general auditing problem, we also consider the "non-constructive auditing" problem as in Definition 7, where the algorithm is only required to tell if there exists an unfair subgroup without returning the unfair certificate. Actually, it turns out any non-constructive auditing algorithm can distinguish the two cases in Theorem 19.

▶ **Corollary 22** (non-constructive auditing is hard). *Given Assumption 11, for any constants $\alpha \in (0,1), \beta \in \mathbb{R}_+$, and any $C/\sqrt{d \log d} \le \epsilon \le c'/\log^{(1+\beta)/2} d$ where C is a sufficiently large constant and c' is a sufficiently small constant, no auditing algorithm can tell if there exists a unfair certificate for halfspace subgroups in $\mathbb{R}^d$ with*

- *an additive error ϵ under Gaussian marginals and running in time $d^{O(1/(\epsilon^2 \log d)^\alpha)}$.*
- *or a multiplicative approximation factor of $1/\mathrm{poly}(d)$ and running in polynomial time.*

Proof. Suppose there exists an auditing algorithm that can either tell if a $\delta\gamma$-unfair certificate or a $\gamma - \epsilon$-unfair certificate exists given a γ-unfair collection of halfspace subgroup and access to data with a Gaussian marginal, where $\delta, \epsilon \in (0,1)$. With the same argument as that of Corollary 20 and 21, we can achieve the desired results. ◀

To the best of our knowledge, there does not exist any PTAS for properly learning general halfspaces in the agnostic model with guarantees of additive error close to $O(1/\sqrt{\log d})$. However, in the next section, we will show that if we restrict out attention to just homogeneous halfspaces under a standard normal distribution, it is possible to achieve additive error of $O(1/\log^{1/C} d)$ for some constant $C > 2$.

5 Auditing Via Agnostic Learning Under Gaussian Distribution

In this section, we present our algorithmic results. Our approach is based on Theorem 12: auditing over subgroups determined by halfspaces can be accomplished by solving a sequence of simpler tasks of learning halfspaces. As a result, we are able to take advantage of existing agnostic learning methods to solve the auditing problem.

Meanwhile, we will discuss the testability of Gaussian distributions and show that existing distribution testing methods [15, 27] for learning halfspaces will not increase the running time significantly for our task. In fact, the running time of the testing method is asymptotically no greater than that of our auditing algorithm.

5.1 Auditing Algorithm for Homogeneous Halfspaces

Assuming there exists an efficient oracle for agnostic learning, Algorithm 1 will eventually return a halfspace h' as a certificate of the subgroup that has the highest unfairness level.

Notice, we create a negatively labelled data sets at Line 3 because maximizing (minimizing) the unfairness $\Pr\{h\}\,|d_{\mathcal{D}}(c,h)|$ for the $c(\mathbf{x}) = 1$ labelling is equivalent to minimizing (maximizing) $\Pr\{h\}\,|d_{\mathcal{D}}(c,h)|$ for $c(\mathbf{x}) = -1$. Thus, by reversing the labels, we can use the oracle to solve both the maximization and minimization directions.

In the loop, we simply follow our previous reduction by dividing the population constraint into multiple approximately-fixed-size constraints at Line 11. Then, we solve each sub-task with a fixed population size by calling the oracle on both data sets at Lines 7 and 8.

Algorithm 1 Fairness Auditing.

Input : $n, a, b, \epsilon, \delta, \mathcal{D}$, classifier c, oracle $\mathcal{O}$
Result: μ', h'
$\hat{\mathcal{X}} \leftarrow$ draw $N(d, \epsilon, \delta)$ i.i.d. samples from $\mathcal{D}$;
$\hat{\mathcal{D}^+} \leftarrow \{\hat{\mathcal{X}}, c(\hat{\mathcal{X}})\}$;
$\hat{\mathcal{D}^-} \leftarrow \{\hat{\mathcal{X}}, -c(\hat{\mathcal{X}})\}$;
$\mu \leftarrow a$;
$(\mu', h') \leftarrow (1, c)$;
while $\mu \leq b$ **do**
  $h_\mu^+ \leftarrow \mathcal{O}(\epsilon, \delta/2n, \mu, \hat{\mathcal{D}^+})$;
  $h_\mu^- \leftarrow \mathcal{O}(\epsilon, \delta/2n, \mu, \hat{\mathcal{D}^-})$;
  if $\left|d_{\mathcal{D}}(c, h_\mu^+)\right| < \left|d_{\mathcal{D}}(c, h_\mu^-)\right|$ **then** $h_\mu^+ \leftarrow h_\mu^-$;
  if $\mu' \left|d_{\mathcal{D}}(c, h')\right| \leq \mu \left|d_{\mathcal{D}}(c, h_\mu^+)\right|$ **then** $(\mu', h') \leftarrow (\mu, h_\mu^+)$;
  $\mu \leftarrow \mu + (b - a)/n$;
end

We give the guarantees of our algorithm below and defer the proof to the appendix.

▶ **Theorem 23** (Auditing Framework). *Given any binary classifier $c : \mathbb{R}^d \to \{-1, +1\}$, a data distribution $\mathcal{D}$ whose 1-dimensional marginals have continuous cumulative distribution functions, and collections of halfspaces $\{\mathcal{H}_\mu^{\mathcal{D}} \mid \mu > 0\}$ over $\mathbb{R}^d$, if there exists an oracle $\mathcal{O}$ that takes $\epsilon, \delta, \mu \in (0, 1)$ and $N(d, \epsilon, \delta)$ labelled i.i.d. samples from $\mathcal{D}$ in the form of $(\mathbf{x}, c(\mathbf{x}))$, runs in time $T(d, \epsilon, \delta)$, and returns a halfspace h_μ such that, with at least $1 - \delta$ probability*

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{h_\mu(\mathbf{x}) \neq c(\mathbf{x})\} \leq \min_{h\in\mathcal{H}_\mu^{\mathcal{D}}} \Pr_{\mathbf{x}\sim\mathcal{D}}\{h(\mathbf{x}) \neq c(\mathbf{x})\} + \epsilon$$

then there exists an algorithm that takes $n \in \mathbb{Z}^+$, $0 < a \leq b < 1$, $\epsilon, \delta \in (0, 1)$ and $O(N(d, \epsilon, \delta/n))$ labeled i.i.d samples from $\mathcal{D}$, runs in time $O(nT(d, \epsilon, \delta/n))$ and returns a halfspace h' as a certificate such that $a \leq \Pr_{\mathbf{x}\sim\mathcal{D}}\{h'\} \leq b$ and

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\{h'\} \left|d_{\mathcal{D}}(c, h')\right| \geq \max_{h\in\mathcal{H}^d} \Pr_{\mathbf{x}\sim\mathcal{D}}\{h\} \left|d_{\mathcal{D}}(c, h)\right| - O(\epsilon)$$

with at least $1 - \delta$ probability.

While our framework heavily relies on the methods of agnostic learning with small additive error, unfortunately, there are no known methods for learning general halfspaces that can achieve additive error better than a constant, even under distributions as nice as standard normal ones.

However, if we restrict our audit to the class of homogeneous halfspaces, Diakonikolas et al. [8] proposed an agnostic learning PTAS for homogeneous halfspaces under Gaussian data. That is, we only audit for subgroups with probability mass 1/2.

▶ **Lemma 24** (Learning Homogeneous Halfspaces [8]). *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(0, \mathrm{I})$. There exists an algorithm that, given $\tau, \epsilon, \delta > 0$, and $N = d^{\mathrm{poly}(1/\tau)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, the algorithm runs in time $\mathrm{poly}(N, d)$, and computes a halfspace $h_{\mathbf{v}}$ such that, with probability at least $1 - \delta$, it holds that $\Pr_{\mathcal{D}}\{y \neq h_{\mathbf{v}}(\mathbf{x})\} \leq (1 + \tau)\min_{h\in\mathcal{H}^N_{1/2}} \Pr_{\mathcal{D}}\{y \neq h(\mathbf{x})\} + \epsilon$.*

Now, notice that Lemma 24 gives us an oracle for auditing halfspace subgroups with population size 1/2 under Gaussian distributions, since by Lemma 15, we know that agnostic learning with fixed threshold will have constant population size under a Gaussian distribution and, hence, is equivalent to auditing with fixed population size. Therefore, we can use this oracle in Algorithm 1 to audit the subgroup class $\mathcal{H}^d_{1/2}$ for $\mathcal{D} = \mathcal{N}(0, \mathrm{I})$. We show our algorithmic guarantee of a PTAS in the following corollary.

▶ **Corollary 25** (Auditing Under Gaussian). *Given any binary classifier $c : \mathbb{R}^d \to \{-1, +1\}$, a data distribution $\mathcal{N}(0, \mathrm{I})$ and a collection of halfspaces $\mathcal{H}^{\mathcal{N}}_{1/2}$ over $\mathbb{R}^d$, there exists an auditing algorithm that takes $\epsilon, \delta > 0$ and $N = d^{\mathrm{poly}(1/\epsilon)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ labeled i.i.d. examples from $\mathcal{N}(0, \mathrm{I})$ in the form of $(\mathbf{x}, c(\mathbf{x}))$, runs in time $\mathrm{poly}(N, d)$, and returns a halfspace h' as a certificate such that $\Pr_{\mathbf{x}\sim\mathcal{D}}\{h'\} = 1/2$ and*

$$|d_{\mathcal{N}}(c, h')| \geq \max_{h \in \mathcal{H}^{\mathcal{N}}_{1/2}} |d_{\mathcal{N}}(c, h)| - 2\epsilon$$

with at least $1 - \delta$ probability.

Proof. We can simply run Algorithm 1 for just one iteration with the same set of parameters except that $\mathcal{D} = \mathcal{N}(0, \mathrm{I})$, $n = 1$, $a = b = 1/2$ and the oracle being as described by Lemma 24 for $\tau = \epsilon$. Notice that Lemma 24 guarantees us that the requirement on the oracle in Theorem 23 is satisfied. Thus, we can refer to the proof of Theorem 23 to establish that running Algorithm 1 for just one iteration suffices. Also, since we only run the algorithm for one iteration, we have $T = 1$, hence, the running time is dominated by the running time of the oracle, which is $\mathrm{poly}(N, d)$. ◀

5.2 Testability Of Gaussian Distribution

Given the assumption that our algorithm only works under Gaussian distributions, one might ask if a set of data examples can be tested to be Gaussian without increasing the running time guarantee in Corollary 25 asymptotically. We will show that this kind of testing can be accomplished within the same running time as our auditing algorithm.

A recent work by Rubinfeld and Vasilyan [27] has proposed a moment matching method for testing Gaussian assumptions specifically for agnostic learning. Their method is based on the observation that linear threshold functions have degree $\mathrm{poly}(1/\epsilon)$ polynomial approximations with additive error of ϵ [19, 8]. Abstractly, this moment matching testing method estimates the moments of the data samples up to degree $O(1/\epsilon^4)$ and check if the element-wise difference between the estimated moments and the actual Gaussian moments are small. They proved that running their testing method along with the agnostic learning algorithm proposed by Kalai et al. [19] will not increase the running asymptotically, i.e., $d^{O(1/\epsilon^4)}$.

To see why the testing method in [27] will not increase the asymptotic running time of our auditing algorithm, we need to dig deeper into the algorithm described by Lemma 24 from [8]. First, they run the learning algorithm of Kalai et al. [19] to get an approximating polynomial of degree $O(1/\epsilon^4)$. Then, they estimate the moments of the outer product of the derivatives of the learned polynomial. Finally, they estimate the classification error of a collection of halfspaces in a subspace of degree $O(1/\epsilon^4)$. See [8] for further details.

The most important observation is that every step in the algorithm stated in Lemma 24 only requires estimating the moments of the data up to degree $O(1/\epsilon^4)$. Thus, running the moment matching testing method of [27] will only require an additional $d^{O(1/\epsilon^4)}$ running time, which will not increase the asymptotic running time of the agnostic learning algorithm in Lemma 24 or our auditing algorithm.

6 Future Work

The major drawback of our result is still the lack of approaches of learning halfspaces with a sub-constant error guarantee for more general distributions. Therefore, a major direction for fairness auditing remains to develop an agnostic learning method with additive error guarantees for broader classes, such as log-concave distributions – subject to the constraints of Corollary 21/Diakonikolas et al. [7]. Even a computationally efficient learning algorithm for general halfspaces that can achieve additive error close to $O(1/\sqrt{\log d})$ under Gaussian distributions would be an interesting improvement.

An alternative direction is to seek stronger guarantees for conjunctions on such families of distributions. Conjunctions are more natural in the context of auditing, and their relative lack of expressive power might enable a better guarantee.

References

1 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.

2 Özgür Akgün, Ian P Gent, Christopher Jefferson, Ian Miguel, and Peter Nightingale. Metamorphic testing of constraint solvers. In *Principles and Practice of Constraint Programming: 24th International Conference, CP 2018, Lille, France, August 27-31, 2018, Proceedings 24*, pages 727–736. Springer, 2018.

3 Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

4 Bart Bogaerts, Stephan Gocht, Ciaran McCreesh, and Jakob Nordström. Certified dominance and symmetry breaking for combinatorial optimisation. *Journal of Artificial Intelligence Research*, 77:1539–1589, 2023.

5 William Cook, Thorsten Koch, Daniel E Steffy, and Kati Wolter. A hybrid branch-and-bound approach for exact rational mixed-integer programming. *Mathematical Programming Computation*, 5(3):305–344, 2013.

6 Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 2013.

7 Ilias Diakonikolas, Daniel Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *International Conference on Machine Learning*, pages 7922–7938. PMLR, 2023.

8 Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Conference on Learning Theory*, pages 1522–1551. PMLR, 2021.

9 Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020.

10 Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pages 5118–5141. PMLR, 2022.

11 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

12 Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

13 Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pages 3417–3426. PMLR, 2021.

14 Xavier Gillard, Pierre Schaus, and Yves Deville. Solvercheck: Declarative testing of constraints. In *Principles and Practice of Constraint Programming: 25th International Conference, CP 2019, Stamford, CT, USA, September 30–October 4, 2019, Proceedings 25*, pages 565–582. Springer, 2019.

15 Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1657–1670, 2023.

16 Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1162–1173. IEEE, 2022.

17 David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

18 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

19 Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

20 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, 2018.

21 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

22 Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

23 Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31, 2018.

24 Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

25 Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf`.

26 Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM (JACM)*, 56(6):1–40, 2009.

27 Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1643–1656, 2023.

28 Robert E Schapire. The strength of weak learnability. *Machine learning*, 5:197–227, 1990.

29 Ji Wang, Ding Lu, Ian Davidson, and Zhaojun Bai. Scalable spectral clustering with group fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 6613–6629. PMLR, 2023.

A Analysis Of Reduction

We formally prove that the auditing program (4) can be approximated by a sequence of simpler optimization problems with arbitrary precision.

Proof of Proposition 14. For conciseness of the proof, we define

$$\alpha(k) := a + \frac{k(b-a)}{n}$$

Since $a \le \Pr\{h^*(\mathbf{x}) = 1\} \le b$ by definition, there must exists a $k \in \{0, \dots, n-1\}$ such that

$$\alpha(k) < \Pr\{h^*(\mathbf{x}) = 1\} < \alpha(k+1)$$

Then, since we assumed that $\mathcal{D}$ has a continuous CDF w.r.t. the normal of h^*, we can construct another halfspace h' by either increasing or decreasing the threshold of h^* until $\Pr\{\mathbf{x} \in h'\}$ hits either $\alpha(k)$ or $\alpha(k+1)$. We thus obtain

$$\begin{aligned}\Pr\{h'(\mathbf{x}) \neq h^*(\mathbf{x})\} =& |\Pr\{h^*\} - \Pr\{h'\}| \\ \le& \alpha(k+1) - \alpha(k) \\ =& \frac{(b-a)}{n}\end{aligned} \tag{10}$$

Let $\mathbf{dom} := \{\mathbf{x} \mid h'(\mathbf{x}) \neq h^*(\mathbf{x})\}$. Then, by the triangle inequality and the fact that $\Pr\{c(\mathbf{x}) = 1\} \le 1$, we have

$$\begin{aligned}|\Pr\{h^*\} d_{\mathcal{D}}(c, h^*)| - \left|\Pr\{h'\} d_{\mathcal{D}}(c, h')\right| \le& |\Pr\{h^*\} - \Pr\{h'\}| + |\Pr\{h' \cap c\} - \Pr\{h^* \cap c\}| \\ \le& \frac{(b-a)}{n} + |\Pr\{h' \cap c \cap \mathbf{dom}\} - \Pr\{h^* \cap c \cap \mathbf{dom}\}| \\ \le& \frac{(b-a)}{n} + |\Pr\{\mathbf{x} \in \mathbf{dom}\}| \\ \le& \frac{2(b-a)}{n}\end{aligned} \tag{11}$$

where the second inequality is obtained by expanding $\Pr\{h \cap c\}$ on the event $\mathbf{x} \in \mathbf{dom}$ using the law of total probability and exploiting the fact that h' always agrees with h^* on the complement of $\mathbf{dom}$, i.e., $\Pr\{h' \cap c \cap \mathbf{dom}^c\} = \Pr\{h^* \cap c \cap \mathbf{dom}^c\}$; the third inequality holds because at most one of $h^*(\mathbf{x}) = 1$ and $h'(\mathbf{x}) = 1$ holds for any $\mathbf{x} \in \mathbf{dom}$ by definition; and the last inequality is due to equation (10).

Finally, due to the optimality of h_k^*, we have

$$\begin{aligned}\Pr\{h_k^*\} |d_{\mathcal{D}}(c, h_k^*)| \ge& \Pr\{h'\} |d_{\mathcal{D}}(c, h')| - \gamma^* + \gamma^* \\ \ge& \gamma^* - \frac{2(b-a)}{n}\end{aligned}$$

by inequality (11) with $\Pr\{h^*(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h^*)| = \gamma^*$. ◀

B Proof Of Hardness

We will need the following proposition from [16, 7] in the proof of theorem 19.

▶ **Proposition 26** ([16, 7] Hardness of cLWE). *Given Assumption 11, for any $d \in \mathbb{N}$, any constants $\kappa \in \mathbb{N}, \alpha \in (0,1), \beta \in \mathbb{R}_+$ and any $\log^\beta d \le k \le Cd$ where $C > 0$ is a sufficiently small universal constant, the problem LWE$(d^{O(k^\alpha)}, \mathcal{N}, \mathbb{S}^{d-1}, \mathcal{N}_\sigma, \mathrm{mod}_T)$ over $\mathbb{R}^d$ with $\sigma \ge k^{-\kappa}$ and $T = 1/C'\sqrt{k \log d}$, where $C' > 0$ is a sufficiently large universal constant, cannot be solved in time $d^{O(k^\alpha)}$ with $d^{-O(k^\alpha)}$ advantage*

The problem of continuous Learning With Error (cLWE) under Gaussian distribution is known to be as hard as LWE. Now we are ready to prove the main theorem.

Proof of Theorem 19. We give an efficient method taking as input samples from a distribution $\mathcal{D}'$, that is either from the alternative hypothesis or the null hypothesis of $\mathrm{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0,\mathrm{I}), \mathbb{S}^{d-1}, \mathcal{N}(0,\sigma), \mathrm{mod}_T)$ from Proposition 26, and generate samples from another distribution $\mathcal{D}$ with the following properties: if $\mathcal{D}'$ is from the alternative (resp. null) hypothesis of the LWE problem, then the resulting distribution $\mathcal{D}$ will satisfy the alternative (resp. null) hypothesis requirement of the theorem for the halfspace auditing problem.

The reduction process can be formulated as follow: for a sample $(\mathbf{x}, y)$ from a instance $\mathcal{D}'$ of the problem $\mathrm{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0,\mathrm{I}), \mathbb{S}^{d-1}, \mathcal{N}(0,\sigma), \mathrm{mod}_T)$ from Proposition 26, we simply output $(\mathbf{x}, c(\mathbf{x})) \sim \mathcal{D}$, where

$$c(\mathbf{x}) = \begin{cases} +1, & \text{if } y \le T/2 \\ -1, & \text{otherwise} \end{cases}$$

We argue that $\mathcal{D}$ satisfies the desired requirement stated above.

For the alternative hypothesis case, let $\mathcal{D}'$ be from the alternative hypothesis case of the LWE. Let $\mathbf{s}$ be the secret vector in the LWE problem. We consider the following two halfspaces:

$$\begin{aligned} h_1(\mathbf{x}) =& \operatorname{sgn}(\mathbf{s}^\top \mathbf{x} - T/6) \\ h_2(\mathbf{x}) =& \operatorname{sgn}(-\mathbf{s}^\top \mathbf{x} + T/3) \end{aligned}$$

If we can show $\left|\Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{h_1(\mathbf{x})=1\} d_\mathcal{D}(c,h_1) + \Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{h_2(\mathbf{x})=1\} d_\mathcal{D}(c,h_2)\right| = \Omega(T)$, then either $h = h_1$ or $h = h_2$ satisfies $\Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{h(\mathbf{x})=1\}\,|d_\mathcal{D}(c,h)| = \Omega(T)$, which implies the desired property of the alternative hypothesis we would like to prove. By Lemma 15, we have

$$\begin{aligned} & 2 \Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{h_1(\mathbf{x})=1\} d_\mathcal{D}(c,h_1) + 2 \Pr_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\{h_2(\mathbf{x})=1\} d_\mathcal{D}(c,h_2) \\ =& \underbrace{\Pr\{\neg c\}(\Pr\{\neg h_1\} + \Pr\{\neg h_2\}) + \Pr\{c\}(\Pr\{h_1\} + \Pr\{h_2\})}_{I_1} \\ & - \underbrace{(\Pr\{c(\mathbf{x}) = h_1(\mathbf{x})\} + \Pr\{c(\mathbf{x}) = h_2(\mathbf{x})\})}_{I_2} \end{aligned}$$

To bound I_1, I_2, we first examine the subset of domain where h_1 and h_2 agree, namely

$$\begin{aligned} B :=& \{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = h_2(\mathbf{x})\} \\ =& \{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = 1 \ \cap \ h_2(\mathbf{x}) = 1\} \\ =& \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{s}^\top \mathbf{x} \in [T/6, T/3]\} \end{aligned}$$

Then, for I_1, by the law of total probability, we have

$$\begin{aligned} I_1 =& \Pr\{c(\mathbf{x}) = -1\}(\Pr\{h_1(\mathbf{x}) = -1\} + \Pr\{h_2(\mathbf{x}) = -1\} + \Pr\{\mathbf{x} \in B\} - \Pr\{\mathbf{x} \in B\}) \\ & + \Pr\{c(\mathbf{x}) = 1\}(\Pr\{h_1(\mathbf{x}) = 1\} + \Pr\{h_2(\mathbf{x}) = 1 \ \cap \ \mathbf{x} \notin B\} + \Pr\{h_2(\mathbf{x}) = 1 \ \cap \ \mathbf{x} \in B\}) \\ \overset{\text{(i)}}{=}& \Pr\{c(\mathbf{x}) = -1\}(1 - \Pr\{\mathbf{x} \in B\}) + \Pr\{c(\mathbf{x}) = 1\}(1 + \Pr\{\mathbf{x} \in B\}) \\ =& 1 + \Pr\{\mathbf{x} \in B\}(\Pr\{c(\mathbf{x}) = 1\} - \Pr\{c(\mathbf{x}) = -1\}) \\ =& 1 + \Pr\{\mathbf{x} \in B\}(2\Pr\{c(\mathbf{x}) = 1\} - 1) \end{aligned}$$

where (i) is because $\{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = -1\}, \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = -1\}, \{\mathbf{x} \in B\}$ are pairwise disjoint and their union equals to $\mathbb{R}^d$, $\{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = 1\}, \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = 1 \cap \mathbf{x} \notin B\}$ are disjoint and their union equals to $\mathbb{R}^d$; and since $\{\mathbf{x} \in B\} \subset \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = 1\}$ by definition, $\{\mathbf{x} \in B\} = \{\mathbf{x} \in B \mid h_2(\mathbf{x}) = 1\}$.

For I_2, because for any $\mathbf{x} \in B$, $h_1(\mathbf{x}) = h_2(\mathbf{x}) = 1$ by construction, and by the law of total probability, we have

$$\begin{aligned}
I_2 =& \Pr\{c(\mathbf{x}) = h_1(\mathbf{x}) \cap \mathbf{x} \notin B\} + \Pr\{c(\mathbf{x}) = h_2(\mathbf{x}) \cap \mathbf{x} \notin B\} + 2\Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} \\
=& \Pr\{\mathbf{x} \notin B\} + 2\Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} \\
=& 1 + \Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} - \Pr\{c(\mathbf{x}) = -1 \cap \mathbf{x} \in B\} \\
=& 1 - \Pr\{\mathbf{x} \in B\}(1 - 2\Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}
\end{aligned}$$

By the definition of c as well as the Alternative case distribution of the LWE problem, $\{\mathbf{x} \in \mathbb{R}^d \mid c(\mathbf{x}) = 1\}$ is equivalent to $\{\mathbf{x} \in \mathbb{R}^d \mid \mod_T(\mathbf{s}^\top \mathbf{x} + z) \leq T/2\}$ for some $z \sim \mathcal{N}(0, \sigma^2)$. Furthermore, we have

$$\{\mathbf{x} \in \mathbb{R}^d \mid \mathrm{mod}_T(\mathbf{s}^\top \mathbf{x} + z) \leq T/2\} \equiv \bigcup_{k \in \mathbb{Z}} \{\mathbf{s}^\top \mathbf{x} + z \in (kT, kT + T/2]\}$$

Notice that $\mathbf{s}^\top \mathbf{x} + z$ is a one dimensional Gaussian random variable, which, by symmetry of Gaussian distribution, implies $\Pr\{c(\mathbf{x}) = 1\} = \Pr\{\cup_{k \in \mathbb{Z}} \{\mathbf{s}^\top \mathbf{x} + z \in (kT, kT + T/2]\}\} = 1/2$. Therefore, combining I_1 and I_2 gives

$$\begin{aligned}
I_1 - I_2 =& 2\Pr\{\mathbf{x} \in B\}(\Pr\{c(\mathbf{x}) = 1\} - \Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}) \\
=& \Omega(T)(1/2 - \Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}) \qquad (12)
\end{aligned}$$

where the last equation is because $\mathbf{s}^\top \mathbf{x} \sim \mathcal{N}(0, 1)$, hence, $\Pr\{\mathbf{x} \in B\} = \Pr\{\mathbf{s}^\top \mathbf{x} \in [T/6, T/3]\} = \Omega(T)$. Since we were only concerned with showing $|I_1 - I_2|$ is large, it suffices to show $\Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\} - 1/2 = \Omega(1)$.

For $\mathbf{x} \in B$, we have $\mathbf{s}^\top \mathbf{x} \in [T/6, T/3]$, therefore $c(\mathbf{x}) = -1$ only if $|z| \geq T/6$. Notice that $z \sim \mathcal{N}(0, \sigma^2)$ and Proposition 26 states that the LWE problem is hard for any fixed constant $\kappa \in \mathbb{N}$ and $\sigma \geq k^{-\kappa}$. Given the constant $\beta \in \mathbb{R}_+$ in this theorem, we can take $\kappa = \lceil 1/2\beta + 1/2 + 1 \rceil$, which is a fixed constant. Then, by Proposition 26, the LWE problem is hard for $\sigma = k^{-\kappa} \leq 1/(k^{3/2}\sqrt{\log d}) = o(T)$. Therefore, by a Gaussian tail bound, we have

$$\Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} \{c(\mathbf{x}) = -1 \mid \mathbf{x} \in B\} \leq \Pr_{z \sim \mathcal{N}(0, \sigma^2)} \{|z| \geq T/6\} = o(1)$$

Plugging the above back into Equation (12), we can conclude that

$$\Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} \{h_1(\mathbf{x}) = 1\} d_\mathcal{D}(c, h_1) + \Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} \{h_2(\mathbf{x}) = 1\} d_\mathcal{D}(c, h_2) = \Omega(T)$$

Thus, either $h = h_1$ or $h = h_2$ must satisfy $\Pr_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}\{h(\mathbf{x}) = 1\} |d_\mathcal{D}(c, h)| = \Omega(T)$, which completes the proof for the alternative hypothesis case.

For the null hypothesis, we can immediately see that $\Pr_{\mathbf{x} \in \mathcal{N}}\{h\} d_\mathcal{N}(c, h) = 0, \forall h \in \mathcal{H}^d$ because $c(\mathbf{x})$ is independent from each $h \in \mathcal{H}^d$.

It remains to verify the time lower bound and the distinguishing advantage for auditing halfspace subgroups. From Proposition 26, we know that under Assumption 11, for the problem $\mathrm{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0, \mathrm{I}), \mathbb{S}^{d-1}, \mathcal{N}(0, \sigma^2), \mathrm{mod}_T)$ with any $\sigma \geq k^{-\kappa}$ (where $\kappa \in \mathbb{N}$ is a

constant) and $T = 1/c'\sqrt{k \log d}$, where $c' > 0$ is a sufficiently large universal constant, the problem cannot be solved in $d^{O(k^\alpha)}$ time with $d^{-O(k^\alpha)}$ advantage. Therefore, under the same assumption, there is no algorithm that can solve the decision version of auditing problem w.r.t. halfspace subgroups in $d^{O(k^\alpha)}$ time with $d^{-O(k^\alpha)}$ advantage. ◀

C Analysis Of Algorithm

We prove the correctness, time and sample complexity of Algorithm 1.

Proof of Theorem 23. Let's notice that, although each iteration of the loop in Algorithm 1 solves $\min_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$ and $\max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$, it is essentially equivalent to solving $\max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} |d_{\mathcal{D}}(c, h)|$ according to Lemma 17. As the oracle returns a halfspace with additive error smaller than ϵ with probability at least $1 - \delta$, we have that

$$\max(\left|d_{\mathcal{D}}(c, h_\mu^+)\right|, \left|d_{\mathcal{D}}(c, h_\mu^-)\right|) \geq \max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \left|d_{\mathcal{D}}(c, h_\mu^+)\right| - \frac{\epsilon}{\mu}$$

with probability at least $1 - \delta/n$ because of Lemma 17 as well as a union bound.

Across all iterations, the algorithm maximizes $\mu \left|d_{\mathcal{D}}(c, h_\mu^+)\right|$ over $\mathcal{H}_\mu^{\mathcal{D}}$ for μ increase from a to b with step size $(b-a)/n$. With a union bound over all n iterations, we obtain the same additive error ϵ in every iteration, with probability at least $1 - \delta$. As a result, the algorithm equivalently solves

$$\begin{aligned} \max_{h \in \mathcal{H}^d} \quad & \Pr_{\mathbf{x} \in \mathcal{D}}\{h(\mathbf{x}) = 1\} \left|d_{\mathcal{D}}(c, h)\right| \\ s.t. \quad & a \leq \Pr_{\mathbf{x} \in \mathcal{D}}\{h(\mathbf{x}) = 1\} \leq b \end{aligned}$$

with probability at least $1 - \delta$ for an additive error at most $2(b-a)/n + \epsilon$ according to Proposition 14, which completes the proof. ◀