Multi-Group Fairness Evaluation via Conditional Value-at-Risk Testing

Lucas Monteiro Paes, Ananda Theertha Suresh, Alex Beutel, Flavio P. Calmon, and Ahmad Beirami

Abstract

Machine learning (ML) models used in prediction and classification tasks may display performance disparities across population groups determined by sensitive attributes (e.g., race, sex, age). We consider the problem of evaluating the performance of a fixed ML model across population groups defined by multiple sensitive attributes (e.g., race **and** sex **and** age). Here, the sample complexity for estimating the worst-case performance gap across groups (e.g., the largest difference in error rates) increases exponentially with the number of group-denoting sensitive attributes. To address this issue, we propose an approach to test for performance disparities based on Conditional Value-at-Risk (CVaR). By allowing a small probabilistic slack on the groups over which a model has approximately equal performance, we show that the sample complexity required for discovering performance violations is reduced exponentially to be at most upper bounded by the square root of the number of groups. As a byproduct of our analysis, when the groups are weighted by a specific prior distribution, we show that Rényi entropy of order $2/3$ of the prior distribution captures the sample complexity of the proposed CVaR test algorithm. Finally, we also show that there exists a non-i.i.d. data collection strategy that results in a sample complexity independent of the number of groups.

Index Terms

multi-group fairness, intersectional fairness, hypothesis testing, intersectionality

I. Introduction

Machine learning (ML) algorithms are increasingly used in domains of consequence such as hiring [1], lending [2], [3], policing [4], and healthcare [5], [6]. These algorithms may display performance disparities across groups determined by legally protected attributes [7], [8] (e.g., sex and race). Discovering and reducing performance disparities across population groups is a recognized desideratum in ML. Over the past decade, hundreds of fairness interventions have been proposed to close performance gaps in a range of prediction and classification tasks while preserving overall accuracy [9]–[16].

While there are many notions of fairness in ML [17], we focus on *group fairness* [13], [18]–[20]. Group fairness is usually concerned with a small number (e.g., two) of pre-determined demographic groups $\mathcal{G}$ and requires a certain performance metric to be approximately equal across the groups in $\mathcal{G}$ — different choices of performance metric lead to different notions of fairness. For example, *equal opportunity* [13], a popular variation of group fairness, requires that false positive rates (FPR) be approximately equal across groups in $\mathcal{G}$. A widely accepted mathematical formalization of group fairness violation is the *largest* gap between a model's performance for a group in $\mathcal{G}$ and its average performance across the entire population. We refer to such worst-case group fairness metrics as *max-gap fairness metrics*.

Multi-group fairness notions aim to increase the granularity of groups in $\mathcal{G}$ for which model performance is measured and compared [21]–[25]. Multi-group fairness metrics answer recent calls for testing for

The work of Lucas Monteiro Paes and Flavio P. Calmon was supported in part by the National Science Foundation under grants CAREER 1845852, CIF 2312667, FAI 2040880, and also in part by a gift from Google.

Lucas Monteiro Paes and Flavio P. Calmon are with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02134, USA (e-mail: lucaspaes@g.harvard.edu; flavio@seas.harvard.edu).

Ananda Theertha Suresh and Ahmad Beirami are with Google Research, New York, NY 10011, USA (e-mail: theertha@google.com; beirami@google.com).

Alex Beutel contributed to this research while at Google Research, New York, NY 10011, USA. He is currently with OpenAI (e-mail: alexb@openai.com) .

performance disparities across groups determined by intersectional demographic attributes [23]–[25] (e.g., groups determined by sex **and** race instead of separate evaluation for each attribute). A fine-grained characterization of groups is critical for discovering performance disparities that may lead to systematic disadvantages across intersecting dimensions [24].

Though socially critical, the need for multi-group fairness and the discovery of max-gap fairness violations pose competing statistical objectives. Intuitively, as the number of groups increases (i.e., $|\mathcal{G}|$ becomes large), so does the sample complexity of evaluating max-gap fairness metrics. In such cases, even testing for multi-group fairness is challenging [21], [22], [26]. For instance, Kearns *et al.* [21] show that a model can be reliably tested for multi-group fairness with a test dataset of size $O(|\mathcal{G}|)$. Testing for multi-group fairness may be infeasible when $|\mathcal{G}|$ is large, such as when $\mathcal{G}$ is a product set of several group-denoting attributes, such as sex, age, and race.

In this paper, we derive fundamental performance limits for testing for multi-group fairness and propose a notion of multi-group fairness based on conditional value-at-risk (CVaR) [27] that allows practitioners to relax fairness guarantees to decrease the sample complexity for reliably testing for multi-group fairness — we call it CVaR fairness. CVaR is a popular metric that is a convex relaxation of the *maximum*, making it a suitable objective for optimization as well. As we shall see, CVaR fairness also leads to a much improved sample complexity as compared to the max-gap fairness allowing us to scale the number of groups in a tractable manner. Remarkably, for non-i.i.d. datasets, we propose an algorithm for reliably testing for CVaR fairness violations with a number of samples independent of the number of demographic groups.

Formally, each individual has sensitive attributes denoted by $(s_1, ..., s_k)$ for some $k \in \mathbb{N}$, and $s_i \in \mathcal{S}^i$ — the combination of which defines a protected group. Intersectionality argues that all the population groups defined by the combinations $(s_1, ..., s_k) \in \mathcal{S}^1 \times ... \times \mathcal{S}^k$ have unique sources of discrimination and privilege [28]. Therefore, it is important to discover and treat disparities in performance across groups $g \in \mathcal{G} = \mathcal{S}^1 \times ... \times \mathcal{S}^k$ [23], [24], [29], [30]. In this case, the number of groups that we aim to ensure that the ML algorithm treats similarly is $|\mathcal{G}| = |\mathcal{S}^1| \cdot |\mathcal{S}^2| ... \cdot |\mathcal{S}^k|$. When each $\mathcal{S}^i$ is at least binary for all $i \in [k]$, $\mathcal{G}$ has an exponential number of groups $|\mathcal{G}| \geq 2^k$.

Let $\mathbf{w} = (w_1, ..., w_{|\mathcal{G}|})$ be a stochastic vector that defines a distribution over groups. Broadly speaking, we can think of $\mathbf{w}$ as a vector of fairness aware importance scores for every group $g \in \mathcal{G}$; for example, $\mathbf{w}$ can be given by the percentage of each group in the population, an importance weight for each group, or uniform $\left(\text{i.e., } w_g = \frac{1}{|\mathcal{G}|}\right)$ when assuming a uniform prior on the group distributions. The theoretical contribution of this work (Theorem 1) makes a connection between the sample complexity for testing multi-group fairness and the Rényi entropy of order $2/3$, which is defined by $H_{2/3}(\mathbf{w}) \triangleq 3 \log_2 \sum_g w_g^{2/3}$. Our **main contributions** are:

- We provide an impossibility result, showing that *any hypothesis test* that uses i.i.d. samples from the data distribution needs at least $n = \Omega(|\mathcal{G}|)$ points to reliably test if max-gap fairness metrics are greater than a fixed threshold — Proposition 1. Our result complements the achievability result in [21], where it was shown that multi-group fairness can be tested with $n = O(|\mathcal{G}|)$ samples. Hence, scaling $\mathcal{G}$ to intersectional groups determined by a combination of features can be infeasible due to the cost of acquiring an exponential number of samples.
- Motivated by the first negative result, we propose a multi-group fairness metric that is a relaxation for max-gap fairness metrics based on the conditional value-at-risk (CVaR) — namely CVaR fairness in Definition 3. We show that it is possible to recover max-gap fairness metrics from CVaR fairness (Proposition 2) and that CVaR fairness lower bounds max-gap fairness metrics (Proposition 3).
- We propose two natural sampling techniques for data collection: *weighted sampling* and *attribute specific sampling* thus providing flexibility for the auditor to collect datasets.
- We propose an algorithm to test if the CVaR fairness metric is greater than a fixed threshold ϵ (Algorithm 1). To do so, we use an estimator for the CVaR fairness metric, use data samples to compute the estimator based on either sampling technique, and perform a threshold test that outputs if the CVaR fairness metric is bigger than the fixed threshold (ϵ) or equal to zero.

- We prove that the proposed Algorithm 1, with weighted sampling, can reliably test for CVaR fairness using at most $n = O\left(2^{\frac{1}{2}H_{2/3}(\mathbf{w})}\right)$ samples, where $H_{2/3}(\cdot)$ is the Rényi entropy of order $\frac{2}{3}$, which is formally defined in (38) (Theorem 1). Our result shows that using Algorithm 1 for the CVaR fairness metric allows us to decrease the sample complexity of testing, achieving a complexity of $2^{\frac{1}{2}H_{2/3}(\mathbf{w})} \leq \sqrt{|\mathcal{G}|} < |\mathcal{G}|$.
- We also show a lower bound for the sample complexity for testing CVaR fairness under weighted sampling when $\mathbf{w}$ is a uniform distribution over all groups. We show that when $\mathbf{w}$ is a uniform distribution over all groups we need $n = \Omega\left(\sqrt{|\mathcal{G}|}\right)$ samples to reliably test for CVaR fairness (Theorem 3). When $\mathbf{w}$ is the uniform distribution, this bound matches the upper bound of the proposed algorithm under weighted sampling tehcnique.
- Finally, we prove that if we have access to 2 samples of any group $g \in \mathcal{G}$ (we will not necessarily use them), it is possible to use Algorithm 1 with attribute specific sampling to reliably test for CVaR fairness utilizing a number of samples (n) that is independent of the number of groups $|\mathcal{G}|$. Specifically, when testing if the CVaR fairness metric is bigger than a threshold ϵ, we only need $O\left(\frac{1}{\epsilon^4}\right)$ samples to test reliably, making the number of samples to test for multi-group fairness independent of the number of groups.

The paper is organized as follows. In Section II, we introduce the notation used in the paper and the hypothesis test framework for auditing a model for multi-group fairness. In Section III, we define the max-gap fairness metric and provide a converse result for auditing a model for max-gap fairness — this result indicates when auditing a model for max-gap fairness is infeasible. In Section IV, we define the conditional value-at-risk fairness metric and describe its properties. In Section V, we define an algorithm along with sampling strategies for efficiently testing whether the CVaR fairness metric is greater than a threshold or is equal to zero. In Section VI we show a converse result for auditing a model for CVaR fairness under certain assumptions, showing that the proposed testing algorithm achieves optimal order, and providing an operational meaning for the Rényi entropy of order $2/3$. Finally, in Section VII we study the practical consequences of our theoretical contributions by (i) empirically analyze the impact of the Rényi entropy on the multi-group fairness auditing accuracy, (ii) computing the sample complexity for reliably auditing using CVaR fairness in comparison with max-gap fairness, and (iii) using our converse bounds to compute the maximun number of groups one model may use for multi-group fairness auditing to be feasible with reasonable performance.

II. Background & Notation

We start by describing each individual as a tuple (x, g, y) where $x \in \mathcal{X}$ is the attribute vector, $g \in \mathcal{G}$ is the *protected* group variable (e.g., $(\texttt{sex, race, age}) = (\texttt{male, asian, 20-30})$), and $y \in \{0, 1\}$ is a binary label to be predicted — note that $x \in \mathcal{X}$ may contain $g \in \mathcal{G}$. Let $\widehat{y} = f_\theta(x) \in \{0, 1\}$ denote a model that aims to predict y and where the model is parameterized by θ, e.g., weights of a neural network. We denote the data of each individual by $z = (x, g, y, \widehat{y}) \in \mathcal{Z}$, when using a random variable instead of the data samples, we write $Z = (X, G, Y, \widehat{Y})$, and assume that the data is drawn from an unknown distribution $\widetilde{P}_Z$. We sample from $\widetilde{P}_Z$ to create a model auditing dataset with $n \in \mathbb{N}$ samples $\mathcal{D}_{\text{audit}} = \{z_i\}_{i=1}^n$ — the data points are not necessarily i.i.d. sampled. Note that the samples $z = (x, g, y, \widehat{y})$ are realizations of the random variable $Z = (X, G, Y, \widehat{Y})$.

We also define $\mathbf{w} \triangleq (w_1, ..., w_{|\mathcal{G}|}) \in \Delta^{|\mathcal{G}|}$ to be a prior distribution on the groups of the audit dataset — the model auditor has full control over $\mathbf{w}$. For example, when testing for fairness, the model auditor may choose $\mathbf{w}$ to be given by a uniform prior, i.e., $w_g = \frac{1}{|\mathcal{G}|}$. This way, all groups would be equally represented in the audit dataset and would receive equal fairness guarantees. The model auditor may choose $\mathbf{w}$ to increase the weight of such groups, reflecting some desired policy.

Group fairness notions such as equal opportunity [13], equal odds [13], and representation disparity [31] are widely used in practice. These notions quantify *fairness* in terms of differences (or lack thereof) of a

given performance metric (e.g., FPR) across population groups. In general, group fairness notions can be represented in terms of a *fairness violation metric*, which will be zero if there is no performance disparity across groups, and positive if some performance disparity exists (e.g., one group has a higher FPR than the rest). We use fairness violation metrics to test if a given machine learning model follows a specific fairness notion. Let $\texttt{F}$ be a fairness violation metric that takes the test dataset distribution $\widetilde{P}_Z$ as input and outputs a real number — $\texttt{F}$ will depend on the group distribution vector $\mathbf{w}$.

We are interested in testing if most groups $g \in \mathcal{G}$ are treated similarly accordingly to the metric (i.e., $\texttt{F}(\widetilde{P}_Z) = 0$) or if there exists a group that is being treated differently than the others (i.e., $\texttt{F}(\widetilde{P}_Z) \geq \epsilon$). Hence, we propose to use a hypothesis test framework for testing if multi-group fairness metrics are bigger than a fixed threshold. This is equivalent to the hypothesis test in Definition 1.

Definition 1 (ϵ-Test): Given a fairness violation metric $\texttt{F}$ for the dataset distribution $\widetilde{P}_Z$ over $\mathcal{Z}$, we call the following hypothesis test an ϵ-test.

$$\begin{cases} \mathrm{H}_0: & \texttt{F}(\widetilde{P}_Z) = 0 \\ \mathrm{H}_1: & \texttt{F}(\widetilde{P}_Z) \geq \epsilon. \end{cases} \tag{1}$$

We denote by $\psi_\epsilon : \mathcal{Z}^n \to \{0, 1\}$ the decision function that takes the dataset $\mathcal{D}_{\text{audit}} = \{z_i\}_{0 \leq i \leq n}$ such that $z_i \in \mathcal{Z}$ and predicts

$$\psi_\epsilon\left(\mathcal{D}_{\text{audit}}\right) = \begin{cases} 0 & \text{if } \mathrm{H}_0 \text{ is true} \\ 1 & \text{if } \mathrm{H}_1 \text{ is true.} \end{cases} \tag{2}$$

The decision function in Definition 1 aims to differentiate between distributions from two separate decision regions ($\mathcal{P}_0$ and $\mathcal{P}_1(\epsilon)$) in the set of all distributions in $\mathcal{Z}$. Specifically, the decision regions $\mathcal{P}_0$ and $\mathcal{P}_1(\epsilon)$ are given respectively by (3) and (4).

$$\mathcal{P}_0 = \left\{ \widetilde{P}_Z \sim \mathcal{Z} \,\middle|\, \texttt{F}(\widetilde{P}_Z) = 0 \right\} \tag{3}$$

$$\mathcal{P}_1(\epsilon) = \left\{ \widetilde{P}_Z \sim \mathcal{Z} \,\middle|\, \texttt{F}(\widetilde{P}_Z) \geq \epsilon \right\} \tag{4}$$

where $\widetilde{P}_Z \sim \mathcal{Z}$ denotes that $\widetilde{P}_Z$ is the dataset probability distribution with support on $\mathcal{Z}$. Assuming a uniform prior on the hypothesis H_0 and H_1, the decision function has a probability of error equal to:

$$P_{\text{error}} = \frac{\Pr(\psi_\epsilon = 1 | \mathrm{H}_0) + \Pr(\psi_\epsilon = 0 | \mathrm{H}_1)}{2}. \tag{5}$$

Since we are distinguishing a single distribution from a set of distributions, the abovementioned test is a composite hypothesis test. Composite hypothesis tests have been recently popularized under *distribution property testing*, see [32] for a detailed survey on the topic.

In this section, we defined fairness violation metrics ($\texttt{F}$) and how to use them to audit a model for multi-group fairness. In the next section, we study a particular case of fairness violation metrics that we call max-gap fairness (Definition 2) and show that performing a reliable ϵ-test for max-gap fairness require at least $O(|\mathcal{G}|/\epsilon^2)$ samples.

III. Max-Gap Fairness

Several fairness violation metrics $\texttt{F}$ measure the magnitude of the difference between model performance for a group and the average performance across all the groups [13], [21], [22], [26], [31]. We call these metrics the max-gap fairness and denote it by $\texttt{F\_MaxGap}$. In this paper, we call the function that measures the performance across different demographic groups the *quality of service* function and the average performance across all groups the *average quality of service*. We give examples of the average quality of service function for equal opportunity in Example 1 and statistical parity in Example 2. We use the quality of service function to formally define the max-gap fairness metrics as follows.

Definition 2 (Max-Gap Fairness): Let $L : \mathcal{Z} \to \{0, 1\}$ be the quality of service function, $\mathbf{w} \in \Delta^{|\mathcal{G}|}$ be the group weight vector, P be a distribution in $\mathcal{Z}$ that measures the average quality of service, and define the average quality of service to be given by

$$\overline{L}(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} w_g \mathbb{E}_P[L(Z)|G = g]. \tag{6}$$

The fairness gap per group is denoted as $\Delta(g; P, L)$ and given by

$$\Delta(g; P, L, \mathbf{w}) \triangleq \left| \mathbb{E}_P[L(Z)|G = g] - \overline{L}(\mathbf{w}) \right|. \tag{7}$$

The *max-gap* fairness metric `F_MaxGap` is the maximum fairness gap per group $\Delta(g; P, L, \mathbf{w})$ for $g \in \mathcal{G}$, i.e.,

$$\texttt{F\_MaxGap}(P, L, \mathbf{w}) \triangleq \max_{g \in \mathcal{G}} \Delta(g; P, L, \mathbf{w}). \tag{8}$$

Note that when L, P, and $\mathbf{w}$ are clear from the context we drop them and just write $\texttt{F\_MaxGap} \triangleq \texttt{F\_MaxGap}(P, L, \mathbf{w})$.

We demonstrate next how group fairness metrics such as Equal Opportunity (EO) and Statistical Parity (SP) can be expressed in terms of the above max-gap fairness definition.

Example 1 (Equal Opportunity): Recall that the equal opportunity violation metric is given by

$$\texttt{F}_{EO} = \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1|Y = 0, G = g) - \Pr(\widehat{Y} = 1|Y = 0) \right| \tag{9}$$

$$= \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1|Y = 0, G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1|Y = 0, G = g)) \right|. \tag{10}$$

Note that we can recover $\texttt{F}_{EO}$ from the max-gap fairness definition by taking

$$L_{EO}(z) = \mathbb{I}_{\widehat{y}=1}(z), \tag{11}$$

$$P_{EO}(z) = \widetilde{P}_Z(z|Y = 0). \tag{12}$$

We have that the fairness gap per group is given by

$$\Delta(g; P_{EO}, L_{EO}, \mathbf{w}) = \left| \Pr(\widehat{Y} = 1|Y = 0, G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1|Y = 0, G = g) \right|. \tag{13}$$

Hence, we recover the equal opportunity violation metric

$$\texttt{F}_{EO} = \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1|Y = 0, G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1|Y = 0, G = g)) \right| \tag{14}$$

$$= \max_{g \in \mathcal{G}} \Delta(g; P_{EO}, L_{EO}, \mathbf{w}) = \texttt{F\_MaxGap}(P_{EO}, L_{EO}, \mathbf{w}). \tag{15}$$

Example 2 (Statistical Parity): Recall that the statistical parity violation metric is given by

$$\texttt{F}_{SP} = \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1|G = g) - \Pr(\widehat{Y} = 1) \right| \tag{16}$$

$$= \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1|G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1|G = g)) \right|. \tag{17}$$

Note that we can recover $\texttt{F}_{SP}$ from the max-gap fairness definition by taking

$$L_{SP}(z) = \mathbb{I}_{\widehat{y}=1}(z), \tag{18}$$

$$P_{SP}(z) = \widetilde{P}_Z(z). \tag{19}$$

We have that the fairness gap per group is given by

$$\Delta(g; P_{SP}(z), L_{SP}, \mathbf{w}) = \left| \Pr(\widehat{Y} = 1 | G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1 | G = g)) \right|. \tag{20}$$

Hence, we recover the statistical parity violation metric

$$\mathtt{F}_{SP} = \max_{g \in \mathcal{G}} \left| \Pr(\widehat{Y} = 1 | G = g) - \sum_{g \in \mathcal{G}} w_g \Pr(\widehat{Y} = 1 | G = g)) \right| \tag{21}$$

$$= \max_{g \in \mathcal{G}} \Delta(g; P_{SP}(z), L_{SP}, \mathbf{w}) = \mathtt{F\_MaxGap}(P_{SP}(z), L_{SP}, \mathbf{w}). \tag{22}$$

Using a similar process, we can recover other fairness metrics such as calibration [33].

Since max-gap fairness measures the violation in fairness metrics, we aim to ensure that the gap $\mathtt{F\_MaxGap}(P, L, \mathbf{w}) \approx 0$. Achieving this condition implies that the model f_θ is *treating* all groups $\mathcal{G}$ in the population similarly per the chosen fairness notion. However, in general, we don't have access to the distribution P; therefore, it is not possible to compute $\mathbb{E}_P[L(Z)|G = g]$ for all $g \in \mathcal{G}$ exactly. Hence, we can't exactly compute $\mathtt{F\_MaxGap}(P, L, \mathbf{w})$ the fairness metric we are interested in. Regardless, we can estimate $\mathtt{F\_MaxGap}(P, L, \mathbf{w})$ using n i.i.d. samples from P. When we are interested in ensuring that the model is treating only two groups similarly, i.e., when $\mathcal{G} = \{0, 1\}$, We can estimate $\mathtt{F\_MaxGap}(P, L, \mathbf{w})$ using $n = O(1/\epsilon^2)$ i.i.d. samples from P and then reliably test if $\mathtt{F\_MaxGap}(P, L, \mathbf{w}) \approx 0$ — this result is an immediate application of Chebyshev inequality [34]. Also, in [21, Theorem 2.11], it was shown that using the same procedure we can estimate $\mathtt{F\_MaxGap}(P, L, \mathbf{w})$ for multi-group fairness using $n = O(|\mathcal{G}|/\epsilon^2)$ i.i.d. samples from P.

In this work, we are interested in the case where $|\mathcal{G}|$ is large — for example, when it contains exponentially many combinations of sensitive attributes (protected groups), and intersectionality is considered. Here, the fairness definitions of interest will be the maximum gap fairness as in (8), but $|\mathcal{G}|$ can grow exponentially. In this regime, estimating the fairness metrics given by $\mathtt{F\_MaxGap}$ (8) is challenging. It is necessary to estimate $\mathbb{E}_P[L(Z)|G = g]$ for all $g \in \mathcal{G}$; hence, if $|\mathcal{G}|$ is sufficiently large, performing all these estimations is impractical [21], [22], [26]. In fact, we show that even *testing* if $\mathtt{F\_MaxGap}$ is above a given threshold is challenging, requiring further assumptions on the hypothesis test procedure to reduce sample complexity.

Next, we show that the max-gap fairness metrics discussed in Definition 2 *require* at least $\Omega(|\mathcal{G}|/\epsilon^2)$ samples for the ϵ-test to have a small probability of error. In Proposition 1, we use the fact that there exist two distributions $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1(\epsilon)$ that are close in the space of distributions in terms of Hellinger distance [35]. Intuitively, the fact that the hypothesis regions are close indicates that it is difficult to distinguish between certain distributions in them. In Proposition 1, we show that the probability of error of the ϵ-test is bounded away from 0.

Proposition 1: (Converse for Max-Gap Fairness). For max-gap fairness metrics $\mathtt{F\_MaxGap}$ (e.g., equal opportunity in Example 1 and statistical parity in Example 2) with quality of service function L and measuring the average quality of service using the distribution P such that $(L, P) \in \{(L_{EO}, P_{EO}), (L_{SP}, P_{SP})\}$. If $\epsilon \in [0, 0.5]$ and $w_g = P(G = g) = \frac{1}{|\mathcal{G}|}$ for all $g \in \mathcal{G}$, using n i.i.d. samples from P the minimum probability of error of performing an ϵ-test for $\mathtt{F\_MaxGap}$ in Definition 1 is lower bounded by

$$2 \inf_{\psi_\epsilon} P_{\text{error}} \geq 1 - \left[2 \left(1 - \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|} \right)^n \right) \right]^{1/2}. \tag{23}$$

Furthermore, it is necessary to have access to n given in (24) i.i.d. samples from P to test if $\mathtt{F\_MaxGap} = 0$ or $\mathtt{F\_MaxGap} \geq \epsilon$ correctly with probability 0.99.

$$n = \Omega\left(\frac{|\mathcal{G}|}{\epsilon^2} \right). \tag{24}$$

We provide the proof for this result in Appendix B.

In Proposition 1, we have shown that the probability of error in the hypothesis test is lower bounded by a function of n the number of samples from $P_{|Y=0}$, $|\mathcal{G}|$ the number of groups, and ϵ and it is necessary to have at least $n = \Omega\left(|\mathcal{G}|/\epsilon^2\right)$ samples to audit a model for max-gap fairness. Thus, a hypothesis test approach cannot decrease the sample complexity of auditing a model for max-gap fairness. Next, we define a fairness metric based on the conditional value-at-risk, aiming to decrease the sample complexity of testing for multi-group fairness, albeit with a weaker notion of fairness violation guarantee.

IV. CVaR Multi-Group Fairness Metrics

In this section, we define the conditional value-at-risk (CVaR) fairness and prove the properties of this multi-group fairness metric.

We adopt the CVaR fairness definition of [36], which is motivated by conditional value-at-risk [27]. Similarly to [36], we are concerned with applying CVaR to the fairness gap $\Delta(g)$. Following [36], Soen *et al.* [37] developed a post-processing technique to decrease CVaR fairness. Meng and Gower [38] applied CVaR to the model loss with the objective of improving worst-case model performance – not applying it to produce a fairer model. Differently from [36]–[38], our main contribution is the definition of CVaR fairness as a relaxation of the max-gap fairness metrics and applying CVaR to quantities beyond the model loss. Additionally, we provide a sample-efficient test for CVaR, which is the main algorithmic contribution of this paper.

The CVaR fairness is an extension of the usual definition of fairness given in Definition 2, i.e., we will show that it is possible to recover Definition 2 from the CVaR fairness in Definition 3.

The definition of CVaR for a random variable W starts by computing $q_\alpha(W)$ the α-quantile of W for $\alpha \in [0, 1)$. The α-quantile of a random variable W is defined to be a number $q_\alpha(W)$ such that

$$\Pr[W \geq q_\alpha(W)] \geq 1 - \alpha \tag{25}$$

$$\Pr[W < q_\alpha(W)] \geq \alpha. \tag{26}$$

Then, the conditional value-at-risk at level α is denoted by $\text{CVaR}_\alpha(W)$ and given by (27) – CVaR was proposed by [27]:

$$\text{CVaR}_\alpha(W) \triangleq \mathbb{E}[W | W \geq q_\alpha(W)]. \tag{27}$$

Intuitively, $\text{CVaR}_\alpha(W)$ measures the tail behavior of W. For example, when $\alpha = 1$ the conditional value-at-risk becomes the maximum over all possible values of W an event that may have vanishing probability. The case of interest in multi-group fairness (when testing is hard) is when only a small group is treated differently from all others, i.e., when $g \in \mathcal{G}$ is small (w_g is small) but $\Delta(g) \geq \epsilon$. Hence, we are interested in the tail behavior of $\Delta(g)$. For this reason, we next define CVaR fairness to capture the desired tail behavior.

Definition 3 (CVaR Fairness): Let $L : \mathcal{Z} \to \{0, 1\}$ be the quality of service function, $\mathbf{w}$ be the group importance weight vector defined by the auditor, and P be the distribution that measures the average quality of service $\overline{L}(\mathbf{w})$. Recall that

$$\Delta(g) = \left|\mathbb{E}_P[L(Z)|G = g] - \overline{L}(\mathbf{w})\right|.$$

For all $\alpha \in [0, 1)$, we denote CVaR fairness as $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w}))$ and define it by

$$Q_\alpha \triangleq \underset{\substack{\sum_{g\in Q} w_g \leq 1-\alpha \\ Q \subset \mathcal{G}}}{\arg\max} \sum_{g \in Q} w_g \Delta(g) \tag{28}$$

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})) \triangleq \frac{1}{1-\alpha} \sum_{g \in Q_\alpha} w_g \Delta(g). \tag{29}$$

We can also write $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w}))$ as

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})) = \frac{1}{1-\alpha} \max_{\substack{\sum_{g\in Q} w_g \leq 1-\alpha \\ Q\subset\mathcal{G}}} \sum_{g\in Q} w_g \Delta(g).$$

When $P, L, \overline{L}(\mathbf{w})$ are clear from the context we denote $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w}))$ by $\texttt{F\_CVaR}_\alpha(\mathbf{w})$. Note that when the group importance weights $\mathbf{w}$ are uniform, then the set Q_α contains $(1-\alpha)|\mathcal{G}|$ groups; hence, the number of groups decreases linearly with alpha, and for $\alpha = (1 - \frac{1}{|\mathcal{G}|})$, the set only contains one of the groups with the highest $\Delta(g)$, thus CVaR recovers max-gap fairness in this case.

Connecting CVaR fairness back to the classical definition of CVaR [27], it can be considered as an application of the classical CVaR for $\Delta(g)$. We claim that, under mild assumptions[1],

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}) = \mathbb{E}[\Delta(g)|\Delta(g) \geq \min_{g\in Q_\alpha} \Delta(g)] = \mathbb{E}[\Delta(g)|\Delta(g) \geq q_\alpha(\Delta(g))] = \textbf{CVaR}_\alpha(\Delta(g)),$$

CVaR fairness captures the behavior of $\Delta(g)$ while accounting for the influence of the group weight on the tail.

Our definition for the CVaR fairness is more lenient than the usual multi-group fairness in Definition 2. Particularly, CVaR fairness gives stronger fairness guarantees when α is close to 1. However, it gives practitioners the flexibility of allowing α to be smaller and make $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ easier to estimate — i.e., decrease the sample complexity to perform an ϵ-test for $\texttt{F\_CVaR}_\alpha(\mathbf{w})$.

Note that when $\alpha = 0$, $\texttt{F\_CVaR}_0(\mathbf{w})$ corresponds to the average fairness gap of all groups. In this case, we can estimate $\texttt{F\_CVaR}_0(\mathbf{w})$ with precision ϵ using $n = O\left(\frac{1}{\epsilon^2}\right)$ samples from P. Thus, it is possible to perform the ϵ-test reliably using $O\left(\frac{1}{\epsilon^2}\right)$ samples from P — we conclude that by a simple application of Chebyshev's inequality [34]. Additionally, α should be chosen to be a fixed constant (i.e., independent of $\mathcal{G}$ and $\mathbf{w}$), we next exemplify the impact of α in the sample complexity of fairness auditing and in Section V-C we show that, tuning α, it is possible to get a more favorable sample complexity for reliable fairness auditing (ϵ-test for CVaR fairness).

As a middle ground example between easy testing ($\alpha = 0$) and the max-gap fairness in Definition 2, consider the following example where the parameter α in CVaR fairness allows trading stronger fairness guarantees (e.g., all groups, including the ones with almost zero probability of occurrence are equally treated) for better sample complexity by increasing the probability mass of the statistics being estimated — i.e., $\max_g \Delta(g)$ may occur in a demographic group with vanishing probability while Example 3 illustrate how CVaR is estimated over a set of mass $(1-\alpha)$.

Example 3 (CVaR Equal Opportunity): Let $\alpha = 0.75$, the quality of service function be L_{EO}, and P_{EO} be given by the equal opportunity fairness metric in Example 1. The fairness gap per group is given by

$$\Delta(g) = \left| \Pr(\hat{Y} = 1|Y = 0, G = g) - \sum_{g\in\mathcal{G}} w_g \Pr(\widehat{Y} = 1|Y = 0, G = g)) \right|. \tag{30}$$

Therefore, the CVaR Equal Opportunity is given by

$$\texttt{F\_CVaR}_{0.75}(\mathbf{w}) = \frac{1}{0.25} \max_{\substack{\sum_{g\in Q} w_g \leq 0.25 \\ Q\subset\mathcal{G}}} \sum_{g\in Q} w_g \Delta(g). \tag{31}$$

Therefore, we need to estimate a quantity with probability mass 0.25, in contrast with Example 1, where estimating a quantity with potentially vanishing probability was necessary. This example highlights the advantage of CVaR fairness in comparison with max-gap fairness, CVaR fairness gives flexibility for the model auditor to select the minimum probability mass $(1-\alpha)$ they can reliably estimate using their limited number of samples.

[1]The mild assumptions are (1) $\sum_{g\in Q_\alpha} w_g = 1 - \alpha$ and (2) if $g \neq g'$ then $\Delta(g) \neq \Delta(g')$. Lemma 11 shows this result.

In Proposition 2, we show that we can recover max-gap fairness (Definition 2) by choosing an appropriate value for α — max-gap fairness is hard to estimate and may demand $O(|\mathcal{G}|)$ samples from P to perform an ϵ-test reliably as saw in Proposition 1.

Proposition 2: ($\texttt{F\_CVaR}_\alpha$ recovers $\texttt{F\_MaxGap}$). For all $\mathbf{w} \in \Delta^{|\mathcal{G}|}$, $L : \mathcal{Z} \to \{0, 1\}$, and any distribution P with support on $\mathcal{Z}$, there exists $\alpha^* \in [0, 1)$ such that

$$\texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) = \texttt{F\_CVaR}_{\alpha^*}(\mathbf{w}; P, L, \overline{L}(\mathbf{w})). \tag{32}$$

When $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$ we have that $\alpha^* = 1 - \frac{1}{|\mathcal{G}|}$.

We show this result in Appendix C. Proposition 2 indicates that by choosing a suitable $\alpha \in [0, 1)$ we can obtain a reasonable trade-off between sample complexity and fairness guarantees. In Sections V-C, we show how to leverage the flexibility provided by α to efficiently test for multi-group fairness.

For all choices of $\alpha \in [0, 1)$ we prove that $\texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w}))$ is lower bounded by $\texttt{F\_CVaR}_\alpha(\mathbf{w})$. This shows why CVaR fairness is a lenient version of the standard multi-group fairness in Definition 2. In Proposition 3, we show that $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ lower bounds $\texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w}))$.

Proposition 3: ($\texttt{F\_CVaR}_\alpha$ is a lower bound for $\texttt{F\_MaxGap}$). For all $L : \mathcal{Z} \to \{0, 1\}$, distribution P with support on $\mathcal{Z}$ and distribution P with support on $\mathcal{Z}$, $\mathbf{w} \in \Delta^{|\mathcal{G}|}$, and $\alpha \in [0, 1)$ we have that

$$0 \leq \texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})) \leq \texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) \leq 1. \tag{33}$$

We provide a proof for this result in Appendix C.

Proposition 3 shows that $\texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) \geq \epsilon$ does not necessarily imply $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$. For this reason, performing an ϵ-test using $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ is more lenient than performing the same test for $\texttt{F\_MaxGap}(L, \overline{L}(\mathbf{w}))$. However, Proposition 3 ensures that if $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$ then $\texttt{F\_MaxGap}(L, \overline{L}(\mathbf{w})) \geq \epsilon$ which is what we will test in Section V-C.

In this section, we have defined the CVaR fairness and showed that (i) it is an extension of the standard multi-group fairness metrics given in Definition 2 (Proposition 2), (ii) CVaR fairness is a more lenient metric than the max-gap fairness metrics in Definition 2, and (iii) CVaR fairness is a value in the square $[0, 1]$ as max-gap fairness (Proposition 3). Moreover, we exemplified that by carefully choosing α we can leverage the trade-off between sample complexity to perform a ϵ-testing and fairness – Example 3 and Proposition 3.

In the next section, we propose a testing algorithm that leverages the trade-off between sample complexity and fairness guarantees from CVaR fairness to efficiently perform an ϵ-test.

V. Testing for CVaR Fairness

A. Data collection process

In order to test for $\texttt{F\_CVaR}_\alpha(\mathbf{w})$, we need to collect an audit dataset $\mathcal{D}_{\text{audit}}$. We propose two ways of collecting this dataset — weighted sampling (Definition 5) and attribute specific sampling (Definition 6). We denote M_g as the number of samples from group g in the audit dataset. In general, M_g can be a random variable.

Due to the cost of acquiring new samples, there is a limit on the dataset size one can collect. For this reason, next, we define the notion of a data budget that a practitioner may have when testing for CVaR fairness.

Definition 4 (Data Budget): A model is audited with a data budget equal to $n \in \mathbb{N}$ when the expected number of samples used to audit the model is equal to n — i.e., (34) holds.

$$n \triangleq \sum_{g \in \mathcal{G}} \mathbb{E}[M_g]. \tag{34}$$

We now propose two ways of collecting data. We start with the weighted sampling, when the audit dataset $\mathcal{D}_{\text{audit}}$ is constructed using i.i.d. samples.

Definition 5 (Weighted Sampling): Let M_g be the number of samples from group $g \in \mathcal{G}$. Under a data budget n, we define the number of samples per group as

$$M_g \triangleq \text{Bin}(n, v_g) \quad \forall g \in \mathcal{G},$$

where $\text{Bin}(n, v_g)$ represents the Binomial distribution with parameters $n \in \mathbb{N}$. Moreover, $\mathbf{v} = (v_1, ..., v_{|\mathcal{G}|}) \in \Delta^{|\mathcal{G}|}$ is the group marginal distribution from the audit dataset $\mathcal{D}_{\text{audit}}$, i.e., $v_g = P_Z(G = g)$, and $\sum_{g\in\mathcal{G}} \mathbb{E}[M_g] = n$.

Note that the group marginal v_g may be different from $\mathbf{w}$ (the underlying or prior group distribution). To provide additional flexibility for model auditor, we set

$$v_g = \frac{w_g^\eta}{\sum_{g'} w_{g'}^\eta}, \tag{35}$$

where $\eta \in [0, \infty)$. Of these, two special cases are $\eta = 0$ (uniform distribution) and $\eta = 1$ (when $v_g = w_g$). In practice, the model auditor selecting the group marginals $\mathbf{v}$ means that they can control the frequency at which a sample from a group is acquired.

The model auditor may also decide to collect the data without ensuring that samples are i.i.d. from some distribution to optimize the sample complexity of performing an ϵ-test. Next, we define attribute specific sampling, a non-i.i.d. sampling strategy that will allow the model auditor to achieve a sample complexity that is independent of the number of groups $|\mathcal{G}|$ — Corollary 2.

Definition 6 (Attribute Specific Sampling): Let M_g be the number of samples from group $g \in \mathcal{G}$, and $\mathbf{w} = (w_1, ..., w_{|\mathcal{G}|})$ be the group weight vector. Under a data budget n, we define the number of samples per group as

$$M_g \triangleq \text{Ber}(\min\{\gamma w_g, 1\})\frac{n}{\gamma} \quad \forall g \in \mathcal{G}.$$

Where $\text{Ber}(\min\{\gamma w_g, 1\})$ represents the Bernoulli distribution with parameter $\min\{\gamma w_g, 1\}$, $\gamma \in [0, +\infty)$, and $\sum_{g\in\mathcal{G}} \mathbb{E}[M_g] = n$.

In the next section we propose an estimator $\widehat{F}(\mathbf{w})$ that we will use to audit the model for CVaR fairness.

B. Proposed estimator

Estimating the value of the CVaR fairness metric ($\texttt{F\_CVaR}_\alpha$) may still be difficult and require a high sample complexity. However, in this work, we are only interested in performing an ϵ-test, i.e., testing if the CVaR fairness metric is bigger than ϵ or equal to zero. For this reason, instead of defining an estimator that tries to approximate $\texttt{F\_CVaR}_\alpha$, we define it with the sole objective of testing if $\texttt{F\_CVaR}_\alpha > \epsilon$. To do so, we introduce the quantities $\widehat{F}_1$ and $\widehat{F}_2$ in Definition 7 that we show are unbiased estimators for $\sum_{g\in\mathcal{G}} w_g \mathbb{E}[L(Z)|G = g]^2$ and $\overline{L}(\mathbf{w})$ respectively in Lemma 5 and Lemma 7. Then, we use these estimators to approximate an upper bound for the CVaR fairness that we show in Lemma 9, as follows

$$\widehat{F} \approx \sum_{g\in\mathcal{G}} w_g \mathbb{E}[L(Z)|G = g]^2 - \overline{L}(\mathbf{w})^2 \geq (1 - \alpha)(\texttt{F\_CVaR}_\alpha(\mathbf{w}))^2.$$

Definition 7 formalizes the definition of the estimator for testing $\texttt{F\_CVaR}_\alpha$.

Definition 7 (Estimator for testing $\texttt{F\_CVaR}_\alpha$): Let $\texttt{F\_CVaR}_\alpha$ be the CVaR fairness metric that uses the quality of service function $L : \mathcal{Z} \to \{0, 1\}$, the average quality of service $\overline{L}(\mathbf{w})$, and uses the distribution P for $\mathcal{Z}$. Define M_g to be a random variable in $\mathbb{N}$, and assume we have access to M_g samples from $P(Z|G = g)$ for each group $g \in \mathcal{G}$ — note that some M_g may be equal to 0. If $M_g = 0$, we define:

$$\frac{\sum_{j=1}^{M_g} L(z_j)}{M_g} \triangleq 0,$$

and if $M_g \in \{0,1\}$ we define

$$\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g} \frac{(\sum_{i=1}^{M_g} L(z_i) - 1)}{M_g - 1} \triangleq 0.$$

Then, we define the estimators $\widehat{F}_1$ and $\widehat{F}_2$ as follows:

$$\widehat{F}_1(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \frac{w_g}{P\left[M_g \geq 2\right]} \frac{\sum_{i=1}^{M_g} L(z_i)}{M_g} \frac{(\sum_{i=1}^{M_g} L(z_i) - 1)}{M_g - 1},$$

$$\widehat{F}_2(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \frac{w_g}{P\left[M_g \geq 1\right]} \frac{\sum_{j=1}^{M_g} L(z_j)}{M_g}.$$

Finally, our CVaR estimator for $\texttt{F\_CVaR}_\alpha$ is defined as:

$$\widehat{F}(\mathbf{w}) \triangleq \widehat{F}_1(\mathbf{w}) - \left(\widehat{F}_2(\mathbf{w})\right)^2. \tag{36}$$

In Appendix D we derive some of the properties, such as the mean and the variance, of the proposed estimator $\widehat{F}(\mathbf{w})$.

The estimator in Definition 7 gives practitioners flexibility by allowing them to sample the number of data points M_g they will use from each group $g \in \mathcal{G}$. Moreover, allowing the number of samples for a group to be zero ($M_g = 0$) may allow a better sample complexity than $O(|\mathcal{G}|)$ — which is why we use this strategy. In the next section we use the proposed estimator along with the introduced sampling strategies to efficiently perform the ϵ-test for the CVaR fairness metric.

In the next section, we propose a specific decision function ψ_ϵ that uses the auditing dataset $\mathcal{D}_{\text{audit}}$ to compute the estimator in Definition 7 and output if the model is CVaR multi-group fair.

C. An Efficient Test For Multi-Group Fairness

In this section, we provide an algorithm to perform the ϵ-test for the CVaR fairness metric. We show that our algorithm can perform ϵ-test reliably with at most $n = O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\epsilon^2(1-\alpha)} + \frac{2^{\frac{1}{3}H_{2/3}(w)}}{\epsilon^4(1-\alpha)^2}\right)$ samples when using weight sampling — where $H_{2/3}$ denotes the Rényi entropy of order $2/3$. Additionally, we show that when using attribute specific sampling, we can perform ϵ-test for CVaR fairness reliably with at most $n = O\left(\frac{1}{\epsilon^4(1-\alpha)^2}\right)$.

We propose Algorithm 1 to perform an ϵ-test by using the estimator for the CVaR fairness metric proposed in Section V-B. The algorithm takes the group importance weights $\mathbf{w}$, the group sampling distributions $\{Q_g\}_{g \in \mathcal{G}}$ (e.g., attribute specific sampling or weighted sampling), and returns which hypothesis is true (i.e., H_0 is true or H_1 is true).

The proposed algorithm starts by defining the number of samples per group M_g using attribute specific or weighted sampling. When the dataset is fixed, M_g is a prefixed constant. However, if the model auditor has control of the dataset construction, they can use weighted sampling (Definition 5) or attribute specific sampling (Definition 6) to define M_g. Next, it computes the CVaR fairness metric estimator $\widehat{F}(\mathbf{w})$ in Definition 7. Finally, it performs a threshold test comparing $\widehat{F}(\mathbf{w})$ with $\frac{(1-\alpha)\epsilon^2}{2}$. We compare $\widehat{F}(\mathbf{w})$ with the threshold $\frac{(1-\alpha)\epsilon^2}{2}$ because in Lemma 9 we show that $\widehat{F} \approx \sum_{g \in \mathcal{G}} w_g \mathbb{E}[L(Z)|G = g]^2 - \overline{L}(\mathbf{w})^2 \geq (1-\alpha)(\texttt{F\_CVaR}_\alpha(\mathbf{w}))^2$. We formalize this process in Algorithm 1.

We show that the sample complexity for an ϵ-test to differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ using Algorithm 1 is more favorable than $O(|\mathcal{G}|/\epsilon^2)$ — differently than the standard multi-group fairness metrics in Definition 2 which needs at least $\Omega(|\mathcal{G}|/\epsilon^2)$ samples to perform a reliable ϵ-test.

Algorithm 1 CVaR Test

Input: $\mathbf{w} \in \Delta^{|\mathcal{G}|}$, $\{Q_g\}_{g\in\mathcal{G}}$
$M_g \sim Q_g$ ▷ Sample M_g s.t. $\Pr(M_g = k) = Q_g(k)$ as in Def. 5, Def. 6, and Thm. 1.
Sample M_g i.i.d. points from $P_{(\mathbf{x},\mathbf{y})|G=g}$ $\forall g \in \mathcal{G}$ ▷ Sample M_g from $P_{(\mathbf{x},\mathbf{y})|G=g}$ for each group.
Compute $\widehat{F}(\mathbf{w})$ as in Def. 7 ▷ Compute using the samples generate in previous step
if $\widehat{F}(\mathbf{w}) \geq \frac{(1-\alpha)\epsilon^2}{2}$ **then**
  return $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$
else if $\widehat{F}(\mathbf{w}) < \frac{(1-\alpha)\epsilon^2}{2}$ **then**
  return $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$
end if

We first prove that the probability of error in (5) is bounded when using Algorithm 1 as the decision function. Theorem 1 shows that the probability of error is bounded when using weighted sampling, and it depends on the test data group distribution $\mathbf{v} = (v_1, ..., v_{|\mathcal{G}|})$.

Theorem 1: (Achievability with weighted sampling). Let $\mathbf{w} \in \Delta^{|\mathcal{G}|}$ be the group weight vector such that $\max_g w_g \leq 1 - \alpha$, $L : \mathcal{Z} \to \{0, 1\}$ be the quality of service function, and $\overline{L}(\mathbf{w})$ is the average quality of service. Suppose we have access to n i.i.d. samples using, i.e., we used weighted sampling in Algorithm 1. Then,

$$Q_g(k) = \Pr(\text{Bin}(n, v_g) = k),$$

where $\mathbf{v} = (v_1, ..., v_{|\mathcal{G}|})$ is the group marginal from P_Z that we sampled i.i.d. to generate $\mathcal{D}_{\text{audit}}$.

The probability of error (P_{error} in (5)) for Algorithm 1 to differentiate $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ from $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ is such that

$$P_{\text{error}} = O\left(\frac{1}{(1-\alpha)^2\epsilon^4 n^2}\sum_g \frac{w_g^2}{v_g^2} + \frac{1}{n(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{v_g}\right). \tag{37}$$

We show this result in Appendix D. Next, we show how large n needs to be to ensure that the upper bound in Theorem 1 is smaller than a fixed $\delta > 0$. This result gives an upper bound for the sample complexity of testing for CVaR fairness using Algorithm 1.

Recall that the Rényi entropy of order ρ for $\rho \geq -1$ is defined as[2]

$$H_\rho(\mathbf{w}) := \frac{1}{1-\rho}\log_2 \sum_{s\in\mathcal{S}} (w_s)^\rho. \tag{38}$$

Notice that $H_1(\mathbf{w}) = -\sum_{g\in\mathcal{G}} w_g \log w_g$ is the Shannon entropy; and $H_0(\mathbf{w}) = |\mathcal{G}|$ is the max-entropy. Corollary 1 gives the upper bound in terms of the Rényi entropy of order $2/3$ of the group distribution $\mathbf{w}$.

Corollary 1: (Sample complexity of Algorithm 1). Under the same assumption of Theorem 1 (using weighted sampling). With probability at least 0.99 [3], comparing $\widehat{F}(\mathbf{w})$ to $(1-\alpha)\epsilon^2/2$, one can differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ using at most the following number of samples

$$n = O\left(\frac{\left(\sum_{g\in\mathcal{G}} \frac{w_g^2}{v_g^2}\right)^{1/2}}{\epsilon^2(1-\alpha)} + \frac{\sum_g \frac{w_g^2}{v_g}}{\epsilon^4(1-\alpha)^2}\right). \tag{39}$$

[2]We extend the definition at $\rho = 1$ via continuous extension.

[3]Because the probability of success in the hypothesis test is not the main focus of our theoretical analysis, we leave the dependence on it to the appendix and only show the results in the main paper for a probability of error of 0.01.

Moreover, when the model auditor can control the choice of group marginal in the test dataset and set $v_g = \frac{w_g^{2/3}}{\sum_{g^*} w_{g^*}^{2/3}}$, it is only necessary to have access to n samples such that

$$n = O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\epsilon^2(1-\alpha)} + \frac{2^{\frac{1}{3}H_{2/3}(w)}}{\epsilon^4(1-\alpha)^2}\right). \tag{40}$$

We show this result in Appendix D. Note that we prove the result for an arbitrary error probability $\delta \in [0, 1]$, but only display the result for $\delta = 0.01$ as the specific probability of error is not key to our analysis. However, we highlight that the dependence on the probability of error δ is not optimal in our bounds and can be automatically improved by using the median of the means technique, improving the dependence to $\log(1/\delta)$ instead of $1/\delta$, as in [39, Page 8].

The sample complexity for the hypothesis test using CVaR fairness in (40) depends on $(1-\alpha)$ and ϵ. When α is close to one, i.e., max-gap fairness is recovered, the sample complexity in (40) might be worse than the sample complexity for the hypothesis-test using max-gap fairness ($n = O(|\mathcal{G}|/\epsilon^2)$) as shown in [21]. This is a consequence of us treating α as a fixed constant bounded away from zero, which allows us to trade off fairness guarantees for a more feasible sample complexity. However, Proposition 2 shows that the sample complexity of CVaR fairness is, at most, the sample complexity of max-gap fairness since for the most strict choice of α, we recover max-gap fairness. When ϵ and $(1-\alpha)$ are constants and bounded away from zero, the dominant term in the sample complexity is the first term $O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\epsilon^2(1-\alpha)}\right)$. Corollary 1 shows that the sample complexity of performing an ϵ-test CVaR fairness using Algorithm 1 with weighted sampling ($v_g \propto w_g^{2/3}$) has a smaller exponent than testing for standard multi-group fairness metrics as shown in Proposition 1 (testing for standard multi-group fairness has sample complexity $O(|\mathcal{G}|)$), because

$$2^{\frac{1}{2}H_{2/3}(\mathbf{w})} \leq 2^{\frac{1}{2}\log(|\mathcal{G}|)} = |\mathcal{G}|^{\frac{1}{2}} \leq |\mathcal{G}|. \tag{41}$$

In the worst-case-scenario, when $\mathbf{w}$ maximizes the Rényi entropy the sample complexity is upper bounded by $O\left(|\mathcal{G}|^{\frac{1}{2}}/((1-\alpha)\epsilon^2)\right)$ providing a smaller sample complexity than i.i.d. testing using standard multi-group fairness metrics as shown in Proposition 1 — Rényi entropy is maximized for $w_g = \frac{1}{|\mathcal{G}|}$ for all $g \in \mathcal{G}$.

In contrast, when ϵ is sufficiently small, the term $O\left(\frac{2^{\frac{1}{3}H_{2/3}(w)}}{\epsilon^4}\right)$ may be dominant. However, this term is still preferable over the original complexity of the problem $O\left(\frac{|\mathcal{G}|}{\epsilon^2}\right)$ when $\epsilon > \frac{1}{|\mathcal{G}|^{1/3}}$. Recall that the case of interest in this paper is when $\mathcal{G}$ contains a large number of groups (e.g., millions of groups) defined by the combination of sensitive attributes, hence the case where $\epsilon > \frac{1}{|\mathcal{G}|^{1/3}}$ is of major interest — here we are not considering constants in the sample complexity.

Moreover, when the model auditor can control the data collection process, they can choose to sample according to Algorithm 1 to decrease the sample complexity further. For example, when the model auditor has access to at least 2 samples of any given group $g \in \mathcal{G}$ in the population and can perform non-i.i.d. sampling, they can use attribute specific sampling. Next, we show that by choosing M_g using attribute specific sampling (Definition 6), we can get a bound on the probability of error for the ϵ-test that is independent of $\mathbf{w}$. We show this result in Theorem 2.

Theorem 2: (Achievability with attribute specific sampling). Let $\mathbf{w}$ be such that $\max_g w_g \leq c \cdot \epsilon^4$ for a sufficiently small constant c and $\max_g w_g \leq 1-\alpha$, $L : \mathcal{Z} \to \{0, 1\}$ the quality of service function, and $\overline{L}(\mathbf{w})$ the average quality of service. Using attribute specific sampling in Algorithm 1, i.e.

$$Q_g(k) = \Pr\left(\text{Ber}(\min(1, \gamma w_g))\frac{n}{\gamma} = k\right),$$

where $\gamma = \frac{n}{2}$.

The probability of error (P_{error} in (5)) for Algorithm 1 to differentiate $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ from $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ is bounded by:

$$P_{\text{error}} = O\left(\frac{1}{\epsilon^4(1-\alpha)^2 n}\right). \tag{42}$$

We show this result in Appendix D. As an immediate consequence of Theorem 2, we can compute the sample complexity of testing using Algorithm 1 with attribute specific sampling. In Corollary 2, we show that using attribute specific sampling allows us to confidently perform an ϵ-test using a number of data points that do not depend on the number of group $|\mathcal{G}|$.

Corollary 2: (Sample complexity of Algorithm 1 using attribute specific sampling). Let $\mathbf{w}$ be such that $\max_g w_g \leq c \cdot \epsilon^4$ for a sufficiently small constant c and $\max_g w_g \leq 1-\alpha$. With probability at least 0.99, comparing $\widehat{F}(\mathbf{w})$ to $(1-\alpha)\epsilon^2/2$, one can differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ using at most the following number of samples n:

$$n = O\left(\frac{1}{\epsilon^4(1-\alpha)^2}\right).$$

We provide the proof for this result in Appendix D — again, we don't show the dependency on the error probability δ, but we prove it in the appendix and provide the bound for $\delta = 0.01$.

In this section, we have shown that it is possible to perform an ϵ-test for CVaR fairness reliably using $O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\epsilon^2(1-\alpha)} + \frac{2^{\frac{1}{3}H_{2/3}(w)}}{\epsilon^4(1-\alpha)^2}\right)$ samples from P when using Algorithm 1 with weighted sampling — which is i.i.d. We also showed that, by using Algorithm 1 with attribute specific sampling (which is non-i.i.d.), we can reliably test for CVaR fairness with $n = O\left(\frac{1}{\epsilon^4(1-\alpha)^2}\right)$ samples from P.

Intuitively, our theoretical results are a consequence of the fact that when the model auditor can control the sampling frequency in specific groups, then they can decide to sample more *uniformly* from all groups. For example, when using weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$, the groups with higher w_g receive a smaller v_g while the groups with smaller w_g receive a higher v_g. Then, if smaller groups receive higher values of $\Delta(g)$, our proposed sampling methods will capture this disparity. On the other hand, the model auditor can decide to use attribute specific sampling, which intuitively first computes $\Delta(g)$ for the groups with higher weights w_g which is a non-i.i.d. sampling strategy wich a sample complexity independent of $|\mathcal{G}|$.

In the next section, we show that by using weighted sampling with $\mathbf{w}$ uniform, it is actually necessary to use $O\left(\frac{2^{\frac{1}{2}H_{2/3}(\mathbf{w})}}{(1-\alpha)^{\frac{1}{2}}\epsilon^2}\right) = O\left(\frac{|\mathcal{G}|^{\frac{1}{2}}}{(1-\alpha)^{\frac{1}{2}}\epsilon^2}\right)$ samples from P for performing an ϵ-test reliably. Indicating that Algorithm 1 has optimal dependency on $\mathbf{w}$ and $|\mathcal{G}|$.

VI. Converse Result for CVaR Fairness Testing under Weighted Sampling

We present next a converse result for the sample complexity of the ϵ-test for CVaR fairness under weighted sampling. Our result ensures that the performance of Algorithm 1 has optimal order when $\mathbf{w}$ is a uniform distribution over all groups. We prove this optimality by showing that no test can differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$ using less than $\Omega\left(\frac{\sqrt{|\mathcal{G}|}}{\epsilon^2}\right)$ i.i.d. samples — Theorem 3. As a consequence of this result, we provide an operational meaning for the Rényi entropy of order $2/3$ as the log of the number of samples that is necessary to perform an ϵ-test for CVaR reliably.

We proved a similar converse result for the max-gap fairness metrics in Proposition 1. For max-gap fairness we were able to design two distributions P_0 and P_1 that are close together in the space of all distributions in $\mathcal{Z}$ according to the Hellinger divergence but at the same time $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_0) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_1) \geq \epsilon$ — Proposition 1. Using these distributions, we leveraged the Le Cam Lemma [40] from the binary hypothesis test literature and showed that it is necessary to have $n = \Omega(|\mathcal{G}|)$ samples to perform an ϵ-test for max-gap fairness reliably — Proposition 1.

Alas, the proof strategy for deriving a converse result for max-gap fairness testing does not directly apply to CVaR testing. The reason is that it is not possible to design distributions P_0 and P_1 as before. In fact, we can show that if two distributions have CVaR fairness far apart, then the total variation distance between them is also far apart. Specifically, if we design $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_0) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_1) \geq \epsilon$ we cannot guarantee that $\text{TV}(P_0||P_1)$ is sufficiently small to match our achievable sample complexity bound. Hence, the approach from Proposition 1 fails.

As an alternative to the approach in Proposition 1, we take a different route and use the Generalized Le Cam Lemma for a mixture of distributions [41]. Instead of showing that there exist distributions P_0 and P_1 with distant CVaR fairness metric but close distance in the space of distributions, we design a sequence of distributions $\{P_u\}_{\mathcal{U}}$ indexed by $\mathcal{U}$ such that $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_u) \geq \epsilon$ for all $u \in \mathcal{U}$, and also P_0 such that $\texttt{F\_CVaR}_\alpha(\mathbf{w}; P_0) = 0$. We then show that the mixture of distributions $\{P_u\}_{\mathcal{U}}$ is close to P_0 according to the χ^2 divergence — Lemma 1.

Recall that the χ^2 divergence between $P \ll Q$ (i.e., P is absolutely continuous with respect to Q) is given by

$$\chi^2(P||Q) \triangleq \mathbb{E}_Q\left[\left(\frac{P(Z)}{Q(Z)} - 1\right)^2\right]. \tag{43}$$

In Lemma 1, we design the mixture distribution $P_u \in \mathcal{P}_1(\epsilon)$ and show that the mixture is close to a distribution $P_0 \in \mathcal{P}_0$ according to the χ^2 divergence.

Lemma 1: [Close Mixture of Distributions] For CVaR fairness metrics $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ with quality of service function L and measuring the average quality of service using the distribution P such that $(L, P) \in \{(L_{EO}, P_{EO}), (L_{SP}, P_{SP})\}$. Consider the hypothesis regions $\mathcal{P}_0$ (3) and $\mathcal{P}_1(\epsilon)$ (4). Assume that $\mathbf{w}$ is a uniform distribution over all groups, i.e., $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$, For all $\alpha \in [0, 1)$ and $\frac{\alpha|\mathcal{G}|^{1/3}}{4} \geq \epsilon > 0$ there exist a sequence of distributions $P_0 \in \mathcal{P}_0$ and $\{P_u\}_{u\in\mathcal{U}} \subset \mathcal{P}_1(\epsilon)$ indexed by elements of the set $\mathcal{U}$ such that if $u \sim \pi$ we have that

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right] \middle|\middle| P_0^n\right) \leq e^{\frac{128(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}} - 1. \tag{44}$$

We prove this result in Appendix E. In Lemma 1, we showed that there exists a mixture of distributions with CVaR fairness greater than ϵ that is close to a distribution P_0 with CVaR fairness equal to 0. Hence, we can use the generalized Le Cam Lemma to compute the minimum number of samples that is necessary to perform an ϵ-test reliably. In Theorem 3 we show that it is necessary to have $n = \Omega\left(\frac{\sqrt{|\mathcal{G}|}}{(1-\alpha)^{1/2}\epsilon^2}\right)$ samples to perform an ϵ-test reliably.

Theorem 3: (Minimum Sample Complexity). For CVaR fairness metrics $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ with quality of service function L and measuring the average quality of service using the distribution P such that $(L, P) \in \{(L_{EO}, P_{EO}), (L_{SP}, P_{SP})\}$. Consider the hypothesis regions $\mathcal{P}_0$ (3) and $\mathcal{P}_1(\epsilon)$ (4). Assume that $\mathbf{w}$ is a uniform distribution over all groups, i.e., $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$. For all $\alpha \in [0, 1)$, $\frac{\alpha|\mathcal{G}|^{1/3}}{4} \geq \epsilon > 0$, the minimum probability of error in the ϵ-test using n i.i.d. samples from P to distinguish between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$ is lower bounded by

$$2 \inf_\psi P_{\text{error}} \geq 1 - \frac{1}{2}\sqrt{e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}} - 1}, \tag{45}$$

then n samples are needed to perform an ϵ-test with with probability of error smaller than 0.01 where n is such that

$$n = \Omega\left(\frac{\alpha^2\sqrt{|\mathcal{G}|}}{(1-\alpha)^{1/2}\epsilon^2}\right). \tag{46}$$

We prove this result in Appendix E — one more time, we provide the dependence on the error probability δ in the appendix, and the result is shown for $\delta = 0.01$. In Theorem 3, we showed that the order of the sample complexity of the Algorithm 1 is optimal when using weighted sampling and $\mathbf{w}$ is uniform. The results in Lemma 1 and Theorem 3 were given in terms of the equal opportunity fairness metric. However, our result can be trivially extended to several fairness metrics, such as statistical parity and equalized odds.

We highlight that the result in Theorem 3 together with Corollary 1 give an operational meaning for the Rényi entropy of order $2/3$ as the log of the number of samples that is necessary to perform an ϵ-test for CVaR reliably.

VII. Numerical Experiments

In this section, we empirically evaluate the performance of the proposed hypothesis test methods within a Bernoulli model. This model allows us to examine test performance by varying three key parameters: (i) the parameter α in Definition 3 (Figure 1), (ii) the entropy of the group distribution $\mathbf{w}$ (Figure 2), and (iii) the number of samples in the hypothesis test (Figure 3).

Additionally, we determine the maximum number of groups for which we can conduct an ϵ-test for both max-gap fairness and CVaR fairness with a probability of error smaller than 45% and a fixed data budget. The maximum number of groups can be derived via the lower bound provided in Proposition 1 and Theorem 3 (Figure 4). For practitioners, this result aids in deciding the number of sensitive attributes (generating groups $g \in \mathcal{G}$) required to ensure reliable auditing for max-gap or CVaR fairness within a fixed data budget. In this section, we take a conservative approach and assume that *Reliable auditing* refers to performing an ϵ-test with a probability of error smaller than 45%; essentially, the test needs to perform slightly better than a random guess to be considered reliable. Our results can also be applied for a smaller probability of error, say δ instead of 0.45, as we show in Appendix D. This impacts our results by a factor of $\frac{1}{\delta}$ instead of $\frac{1}{0.45}$, decreasing the number of possible groups for reliable testing — this dependence can be improved to $\log\left(\frac{1}{\delta}\right)$ by using the median of the means technique from [39, Page 8].

A. Experimental Setup

First, we define the dataset distribution ($\widetilde{P}_Z$) and the quality of service we aim to ensure equality across different groups in the population. Our experiment aims to ensure statistical parity across groups (Example 2) using max-gap and CVaR fairness.

We start this section by defining the protected groups $\mathcal{G}$.

a) The Demographic Groups: We use $d \in \mathbb{N}$ binary sensitive attributes — in this experiment, we don't specify such groups; however, they can encode information about sex, HIV status, or even the occurrence of stroke. Hence, each demographic group is given by a binary sequence of length d; particularly, we define the set of all protected groups as all elements of $\mathcal{G}(d) = \{0,1\}^d$ — e.g., when $d = 2$ the protected groups are given by $\mathcal{G}(2) = \{(0,0);(1,0);(0,1);(1,1)\}$. We denote the groups by $g \in \mathcal{G}(d) = \{0,1\}^d$ and g^i indicates the i-th entry of the group denoting vector in $\mathcal{G}(d)$.

We define the probability of occurrence of each group $g \in \mathcal{G}(d)$ as the probability of the product of d Bernoulli distributions with parameter $p \in [0,1]$ to be equal to the group sensitive attributes. Rigorously, we define $\mathbf{w} = (w_1, w_2, ..., w_{2^d})$ as:

$$w_g \triangleq \prod_{i=1}^{d} \Pr\left(\text{Ber}(p) = g^i\right). \tag{47}$$

We define the group weights as in (47) for the entropy of $\mathbf{w}$ to be equal to d times the entropy of the Bernoulli random variable, specifically $H_{2/3}(\mathbf{w}) = dH_{2/3}((p, 1-p))$. In Figures 2 and 3, we vary the parameter p to analyze the impact of the entropy of the group weight vector $\mathbf{w}$ on the hypothesis test performance.

Now that we have defined the groups considered in our numerical analysis, we introduce the prediction dataset distribution, i.e., the distribution for $z = (g, \hat{y})$

b) Data Distribution: Each individual is represented by a vector $z = (g, \hat{y})$ where g is its group membership and $\hat{y} \in \{0, 1\}$ is the outcome we aim to ensure statistical parity. The outcome distribution is given by

$$\Pr\left(\hat{y} = 1 | G = g\right) \triangleq q_g, \tag{48}$$

where $q_g \in [0, 1]$ is the per group quality of service parameter and $\hat{y}_{|G=g}$ is a Bernoulli random variable with parameter q_g. Hence, the dataset distribution of each data point is given by

$$\widetilde{P}_Z\left((G, \widehat{Y}) = (g, \hat{y})\right) = w_g q_g. \tag{49}$$

Now that we have defined the groups, we aim to ensure equal treatment across all elements of $\mathcal{G}(d)$. We recall the quality service function and distribution, respectively, $L_{SP}(z)$ and $P_{SP}(z)$, used for measuring the quality of the service in each group and compute the Max-Gap and CVaR fairness for the model.

c) Max-Gap and CVaR Fairness: The quality of service function and the measure for the quality of service of interest are, respectively, L_{SP} and P_{SO} given by

$$L_{SP}(g, \hat{y}) = \mathbb{I}_{\hat{y}=1}(g, \hat{y}) \tag{50}$$

$$P_{SP}(g, \hat{y}) = \widetilde{P}_Z(g, \hat{y}). \tag{51}$$

Therefore, the fairness gap per group $g \in \mathcal{G}$ is given by $\Delta(g)$ as follows

$$\Delta(g) = \left| q_g - \sum_{g \in \mathcal{G}} q_g w_g \right|, \tag{52}$$

which implies that the Max-Gap and CVaR fairness are given by

$$\texttt{F\_MaxGap} = \max_{g \in \mathcal{G}} \Delta(g), \tag{53}$$

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}) = \frac{1}{1-\alpha} \max_{\substack{\sum_{g \in Q} w_g \leq 1-\alpha \\ Q \subset \mathcal{G}}} \sum_{g \in Q} w_g \Delta(g). \tag{54}$$

d) Hypothesis Testing: We aim to test if all groups are treated similarly ($\texttt{F\_MaxGap} = 0$ or $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$) or if there is a group that is unfairly treated ($\texttt{F\_MaxGap} \geq \epsilon$ or $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$), i.e., we aim to perform an ϵ-test. However, in general, we assume we don't have access to the data distribution; hence, to simulate this scenario, we perform the ϵ-test using samples using (i) weighted sampling and (ii) attribute specific sampling. We perform the *hypothesis test for CVaR fairness* using the Algorithm 1 that we propose. We use weighted sampling (Definition 5) with $\mathbf{v} = \mathbf{w}$ and $\mathbf{v} = \mathbf{w}^{2/3}$, and attribute specific sampling. We adopt a *baseline testing for max-gap fairness* that compares the empirical average performance with the empirical average per group. Specifically, when n samples are available $\{z_i\}_{i=1}^n$ each group has M_g samples such that $\sum_g M_g = n$, the baseline method outputs that $\texttt{F\_MaxGap} \geq \epsilon$ when

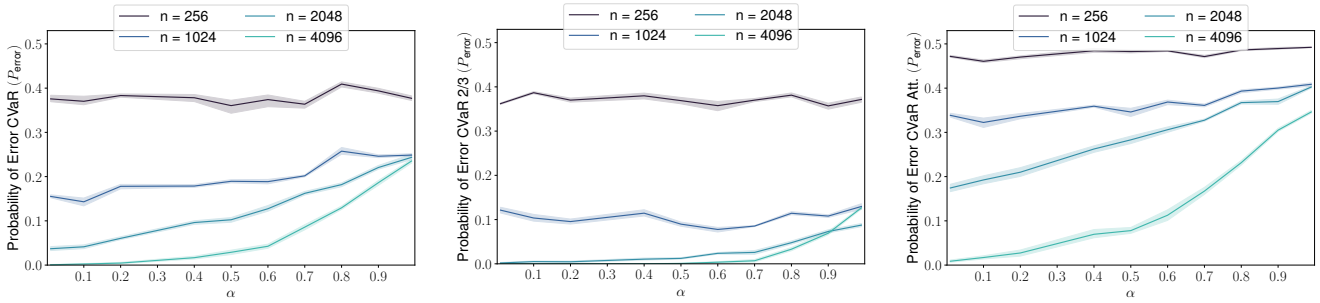
$$\max_{\substack{g \in \mathcal{G} \\ M_g > 0}} \left| \frac{\sum_{i=1}^{M_g} L_{SP}(z_i)}{M_g} - \frac{\sum_{g \in \mathcal{G}} \sum_{i=1}^{M_g} L_{SP}(z_i)}{n} \right| \geq \epsilon, \tag{55}$$

and outputs that $\texttt{F\_MaxGap} = 0$ otherwise. When performing the baseline test for max-gap fairness, we sample i.i.d. from the data distribution that is equivalent to weighted sampling with $\mathbf{v} = \mathbf{w}$.

e) Disparate Quality of Service: To generate disparate quality of service across groups, i.e., $\Delta(g) > 0$ for some $g \in \mathcal{G}$ we take the following approach.

Approach 1: We set the group quality parameter $q_g = 0.05$ for 20% of groups that we choose uniformly at random and $q_g = 0.5$ for the other groups. Specifically, we sample $\lfloor 0.2|\mathcal{G}| \rfloor$ groups $\{g_1, ..., g_{\lfloor 0.2|\mathcal{G}| \rfloor}\}$ and define $q_g = 0.05$ for $g \in \{g_1, ..., g_{\lfloor 0.2|\mathcal{G}| \rfloor}\}$ and $q_g = 0.5$ for all other groups.

Now that we have laid the background for our experiments, we will show our numerical results in the next section.



(a) $\mathbf{w}$ Weighted (b) $\mathbf{w}^{2/3}$ Weighted (c) Attribute Specific

Fig. 1: Probability of error in ϵ-test vs. α for different number of sample sizes. We use $\epsilon = 0.4\times\texttt{F\_CVaR}_\alpha(\mathbf{w})$ in Algorithm 1 — we exactly compute CVaR fairness to evaluate the probability of error. We use $d = 9$ sensitive attributes leading to 512 groups, and the group distribution parameter is taken to be $p = 0.2$. The probability of error was computed using 1000 realizations with the Monte Carlo method, and confidence intervals are plotted using Bootstrap from Seaborn [42] with 95% confidence.

B. Experimental Results

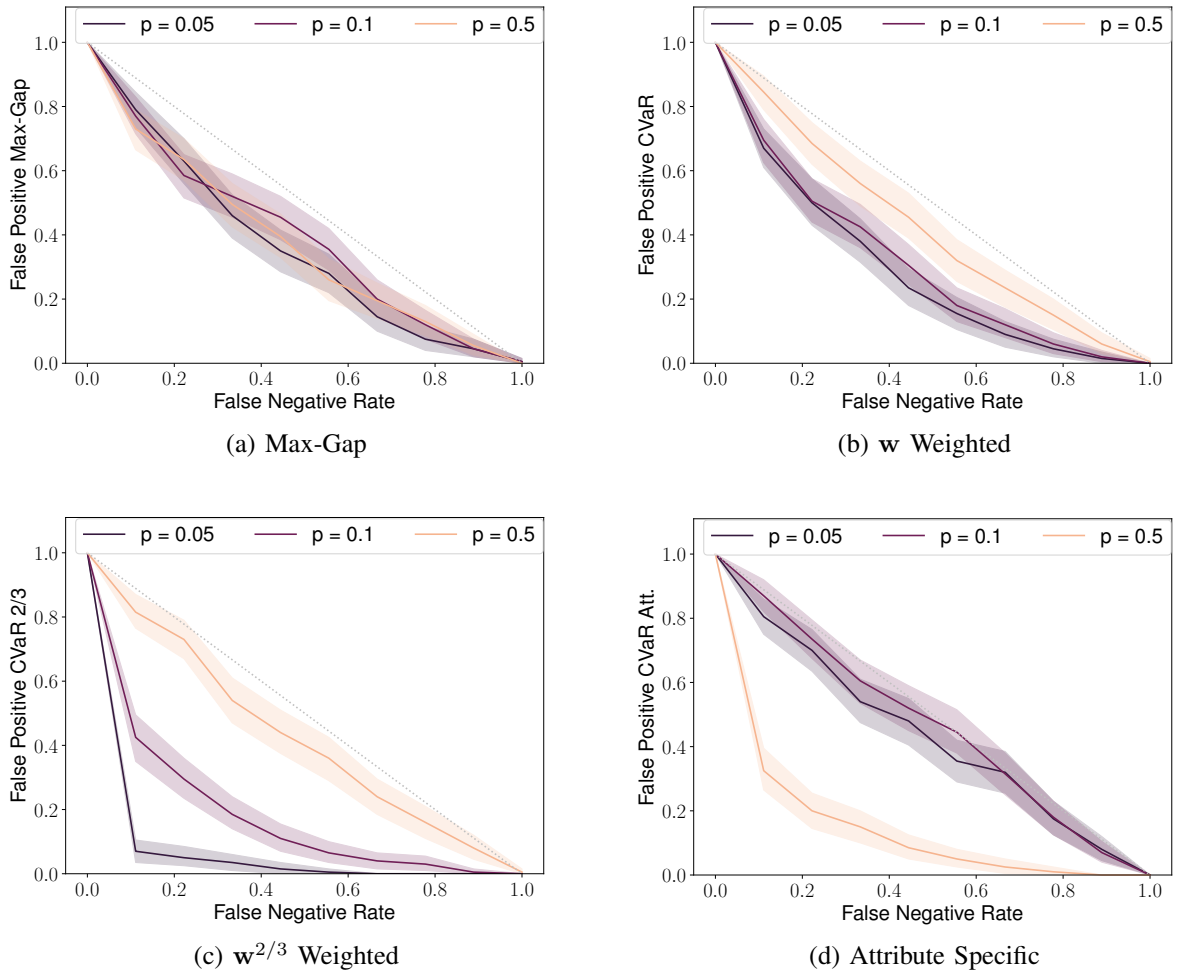
a) On the dependence on α: Figure 1 shows how the probability of error in the ϵ-test for CVaR fairness (in terms if error probability (5)) using Algorithm 1 depends on the choice of the α parameter in CVaR fairness Definition 3. The per group quality of service was defined using Approach 1, i.e., we set 20% of all groups chosen uniformly at random to received a quality of service equals to $q_g = 0.05$ while the other groups received $q_g = 0.5$.

Figure 1 indicates that when α increases, the probability of error when performing an ϵ-test also increases. However, in the low data regime, this behavior changes. In this case, the errors that stem from the limited sample size dominate the probability of error, as we can conclude from the unchanged probability of error when $n = 256$ across both weighted sampling strategies and also attribute specific sampling Figures 1 (a), (b), and (c). The error dominance from the small sample size $n = 256$ is even higher in attribute specific sampling (Figures 1 (c)).

Additionally, we also observe that the dependency on α is more severe in attribute-specific sampling in comparison with weighted sampling using $\mathbf{v} \propto \mathbf{w}^{2/3}$ or $\mathbf{v} = \mathbf{w}$ — which was expected from our error probability bound. Corollary 1 shows that the number of samples necessary to reliably perform an ϵ-test for CVaR fairness using weighted sampling depended on $\frac{1}{(1-\alpha)}$, while Corollary 2 shows that the number of samples depends on $\frac{1}{(1-\alpha)^2}$ when using attribute specific sampling, implying that α has a more severe impact on the error probability of attribute specific sampling when compared with weighted sampling.

b) False positive vs. false negative rate for different values of entropy: Figure 2 shows how the false positive rate depends on the false negative rate for different entropy values achieved by varying the group distribution parameter $p \in \{0.05, 0.1, 0.5\}$). To compute the false positive rate at a given false negative rate, we tune ϵ in Algorithm 1 such that the false negative rate is given by the x-axis and use this ϵ to test if $\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon$. We select the threshold in the baseline test for max-gap (55) similarly to test if $\texttt{F\_MaxGap} \geq \epsilon$.

Figure 2 indicates that weighted sampling for CVaR testing has a more favorable trade-off between false negatives and positives when compared with max-gap fairness testing, specifically for lower entropy values. Notably, the attribute specific sampling method (Figure 2 (d)) exhibits inferior or comparable performance compared to weighted sampling with group weights $\mathbf{v} = \mathbf{w}$ (Figure 2 (b)) or $\mathbf{v} \propto \mathbf{w}^{2/3}$ (Figure 2 (c)) across most entropy values obtained by varying the parameter p. However, under the condition of uniformly distributed group weights (i.e., $p = 0.5$), where entropy is maximized, attribute specific sampling yields the best trade-off between false negative and positive rates, indicating that attribute specific sampling has a better performance when the entropy $\mathbf{w}$ is higher — this result is also supported by Figure 3.



(a) Max-Gap

(b) $\mathbf{w}$ Weighted

(c) $\mathbf{w}^{2/3}$ Weighted

(d) Attribute Specific

Fig. 2: False positive rate at a given false negative rate for the hypothesis tests described in the paper. We use $d = 10$ sensitive groups, generating 1024 groups and under a data budget of $n = 512$ Bernoulli realizations. We use Approach 1 to distribute the per group quality of services. The false positive rate was computed using 1000 realizations in the Monte Carlo method, and confidence intervals are plotted using Bootstrap from Seaborn [42] with 95% confidence.

Weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ (Figure 2 (c)) has the most favorable trade-off between false negative and positive rates for smaller entropies of $\mathbf{w}$ — as is suggested by the direct dependence of its sample complexity on the $H_{2/3}(\mathbf{w})$ as we shown in Corollary 1. At maximum entropy ($p = 0.5$), weighted sampling and the baseline test for max-gap fairness exhibit similar performance, but attribute specific sampling clearly outperforms both weighted sampling and the baseline test for max-gap fairness.

Remarkably, weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ achieves false positive rates smaller than 10% at all false negative rates larger than 20% when $p = 0.05$. Therefore, we conclude that weighted sampling and attribute specific sampling, when used for CVaR fairness testing, exhibit a better or comparable false negative vs. positive trade-off than the baseline method for max-gap fairness testing (55) across all tested values for the parameter p that controls the group weights $\mathbf{w}$ entropy. Moreover, we observe that weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ is more flexible, achieving reasonable performance for different $\mathbf{w}$ entropy values; however, its performance is degraded when entropy is maximized. Attribute specific sampling complements weighted sampling by achieving a more favorable when the entropy of $\mathbf{w}$ is maximized.

c) The dependence on number of samples: Figure 3 shows how the area under the false negative versus positive curve (AUC) changes as the number of samples increases for the ϵ-test in two scenarios

with varying entropy for $\mathbf{w}$. First, Figure 3 shows the AUC for the baseline method (55) when performing an ϵ-test for max-gap fairness (Figure 3 (a)). Second, it shows the AUC when performing an ϵ-test for CVaR fairness using Algorithm 1 with weighted sampling with $\mathbf{v} = \mathbf{w}$ (Figure 3 (b)), weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ (Figure 3 (c)), and attribute-specific sampling (Figure 3 (d)). The per group quality of service is distributed using Approach 1.

First, we note that Figure 3 shows that the performance of the baseline method for max-gap testing doesn't achieve reasonable performance across all entropy values and data budgets tested, i.e., AUC is always larger than 0.3 for max-gap testing. Meanwhile, when using Algorithm 1 with any of the proposed sampling methods for CVaR testing, we observe that their AUC is smaller than 0.2 with just $n = 300$ samples, indicating the clear advantage of CVaR fairness testing in comparison with max-gap fairness.

Additionally, Figure 3 shows that AUC decreases faster for CVaR fairness testing when using weighted sampling in comparison with the attribute specific sampling or the baseline method for max-gap fairness testing in most of the tested scenarios. As we also observed in Figure 1, weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ has a flexible performance, i.e., it performs well under differentvalues of the entropy of $\mathbf{w}$ and data budgets (n).

Figure 3 also shows that using Algorithm 1 with attribute specific sampling has almost perfect performance (AUC ≈ 0) when the groups are uniformly distributed, i.e., $p = 0.5$, even with $n = 100$ samples while testing for CVaR fairness across 1024 groups. When the groups are uniformly distributed, no other tested method achieves such AUC value even with $n = 1500$ samples. As we also observed from Figure 1, these results indicates that when $\mathbf{w}$ is close to uniform CVaR testing using attribute specific sampling offers the best performance across all tested groups while weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ is the best-performing sampling method when the entropy of $\mathbf{w}$ is small. In all cases, the baseline method (55) for max-gap fairness testing doesn't achieve reasonable performance.

d) Maximum $\mathcal{G}$ for Reliably Auditing: Figure 4 uses the lower bound on the error probability from Proposition 1 and Theorem 3. In both cases, we first showed a lower bound for the minimum probability of error in a ϵ-test that uses i.i.d. samples from the data distribution; then we computed the maximum number of samples to make the lower bound smaller than a fixed error probability. Here, we take a different approach, fixed an ϵ threshold and the number of samples used to perform the ϵ-test; we found the maximum number of groups that can be used while ensuring that there exists a decision function that achieves at least an error probability of 45% (almost a random guess). Specifically, we denote the maximum number of groups that can be used for reliably (error probability smaller than 45%) test for max-gap fairness as $G_{\texttt{max-gap}}(n, \epsilon)$ and for CVaR fairness as $G_{\texttt{CVaR}}(n, \epsilon, \alpha)$ and define then next.

In Proposition 1, we showed that the minimum probability of error is lower bounded by (56), and if we take it to be 0.45 we have that the number of groups $|\mathcal{G}|$ is such that

$$2 \inf_{\psi_\epsilon} P_{\text{error}} = 0.9 \geq 1 - \left[2 \left(1 - \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|} \right)^n \right) \right]^{1/2}. \tag{56}$$

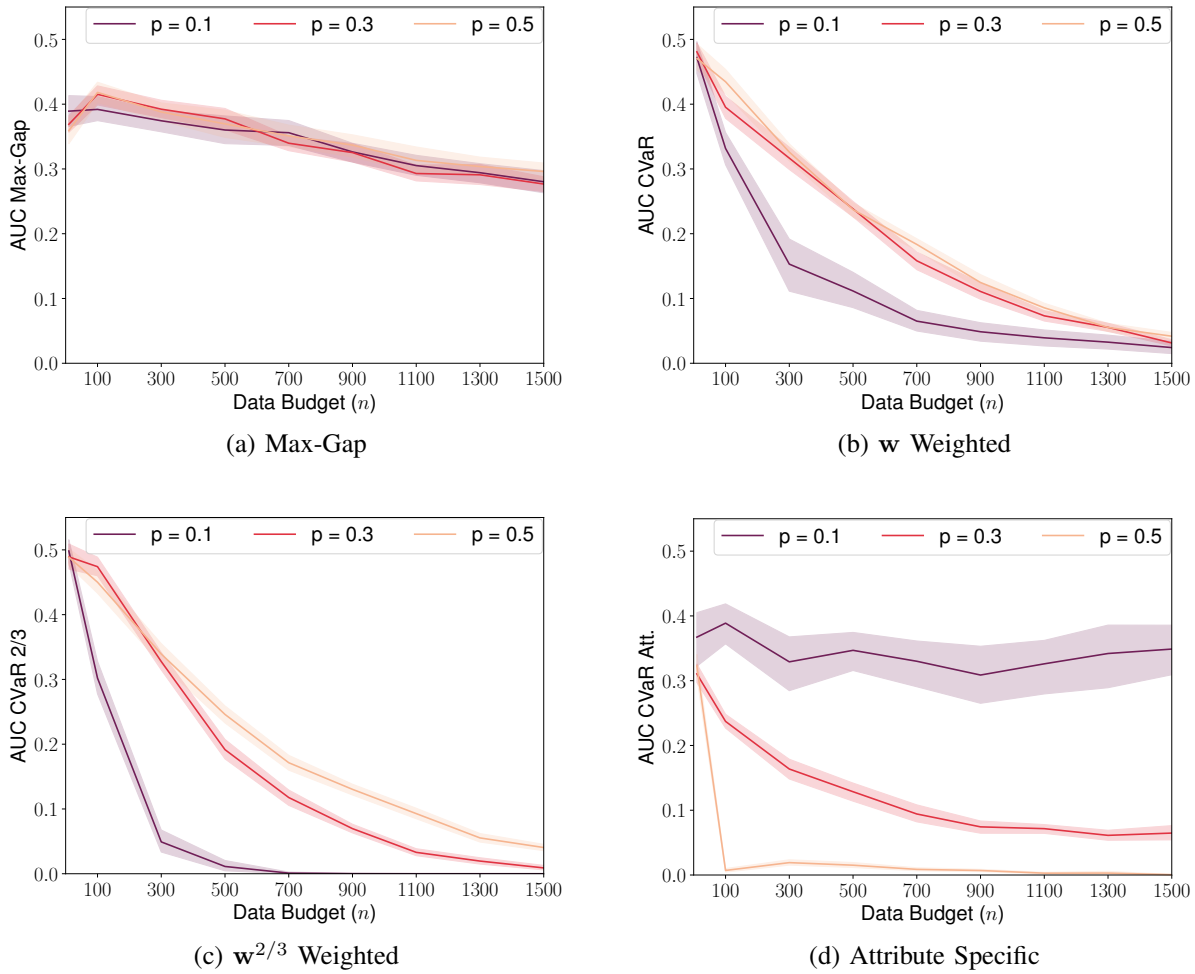
Then, the maximum number of groups we can have while ensuring that the probability of error in testing for max-gap fairness is at most 45% is given by

$$G_{\texttt{max-gap}}(n, \epsilon) \triangleq \arg\max_{|\mathcal{G}|} \left\{ |\mathcal{G}| \in \mathbb{N} \middle| 0.9 \geq 1 - \left[2 \left(1 - \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|} \right)^n \right) \right]^{1/2} \right\}. \tag{57}$$

Similarly, using the lower bound for the probability of error when performing an ϵ-test for CVaR fairness proved in Theorem 3, the maximum number of groups to ensure that the probability of error is at most 45% is given by

$$G_{\texttt{CVaR}}(n, \epsilon, \alpha) \triangleq \arg\max_{|\mathcal{G}|} \left\{ |\mathcal{G}| \in \mathbb{N} \middle| 0.9 \geq 1 - \frac{1}{2} \sqrt{e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4 |\mathcal{G}|}} - 1} \right\}. \tag{58}$$

We show $G_{\texttt{max-gap}}(n, \epsilon)$ and $G_{\texttt{CVaR}}(n, \epsilon, \alpha)$ in Figure 4.



(a) Max-Gap

(b) $\mathbf{w}$ Weighted

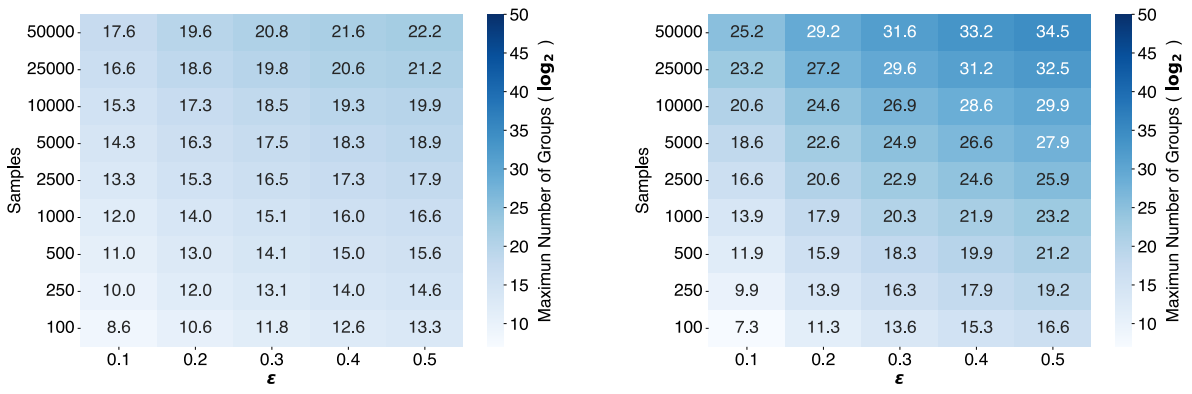
(c) $\mathbf{w}^{2/3}$ Weighted

(d) Attribute Specific

Fig. 3: Area under the false negative vs. false positive curve (AUC) versus the number of samples used to perform the ϵ-test. We vary the entropy by changing the group distribution Bernoulli parameter p with a fixed number of sensitive attributes $d = 10$, generating 1024 groups and data budget in x-axis. We use 100 samples in the Monte Carlo method and repeat this process 20 times to estimate AUC and plot confidence intervals using Bootstrap from Seaborn [42] with 95% confidence.

Remarkably, when $50k$ samples are available, the maximum number of groups a model can have while it is still possible to perform an 0.1-test using max-gap fairness with a probability of error of at most 45% (almost a random guess) is $\lfloor 2^{17.6} \rfloor$ (Figure 4 (a)) while, under the same scenario, it is possible to perform an 0.1-test using CVaR fairness with $\lfloor 2^{25.2} \rfloor$ using $\alpha = 0.9$ (Figure 4 (b)). In practice, this result implies that it is impossible (without further assumption on the dataset distribution) to determine if a machine learning model that was deployed using 18 binary sensitive attributes has a max-gap fairness bigger than 0.1. On the other hand, using CVaR fairness testing, it is possible to reliably test if a machine-learning model that was deployed using 25 binary sensitive attributes has CVaR fairness bigger than 0.1 when we set $\alpha = 0.9$ (the number of sensitive attributes grows with α). Therefore, our bounds indicate that CVaR fairness allows practitioners to deploy machine learning models with more sensitive attributes while reliably ensuring that the model is fair according to CVaR fairness.

e) Concluding remarks from experiments: The experiments indicate that testing for CVaR fairness has a more favorable trade-off between sample complexity and the number of demographic groups in comparison with max-gap fairness, as our theoretical results demonstrate. Therefore, when the data budget to audit a model for fairness is not large enough (smaller than the number of groups) for max-gap fairness



(a) Max-Gap

(b) CVaR ($\alpha = 0.9$)

Fig. 4: The maximum number of groups $\mathcal{G}$ (z-axis) for each choice of threshold ϵ (x-axis) and sample size (y-axis) to ensure that the probability of error in the hypothesis test from Definition 1 is smaller than 45%. Figure (a) shows the maximum number of groups when testing if max-gap fairness is bigger than ϵ. Figures (b) the maximum number of groups when testing if CVaR fairness is bigger than ϵ using an α value of 0.9. The maximum number of groups was computed using (57) for max-gap and (58) for CVaR fairness.

testing, we advocate for practitioners to use CVaR fairness with the largest possible α that still maintains the sample complexity under their data budget — result showed in Figure 4. Moreover, when the group weight vector $\mathbf{w}$ has high Rényi entropy, our results indicate that attribute specific sampling is more efficient in testing for CVaR fairness — as our theoretical results show given that the sample complexity of attribute specific does not increase with the entropy of $\mathbf{w}$. When the entropy of $\mathbf{w}$ is not too large, using weighted sampling with $\mathbf{v} \propto \mathbf{w}^{2/3}$ yields an accurate ϵ-test even under a small data budget as shown in Figure 3 — this is also expected from our theoretical results in Theorem 1. Finally, when only i.i.d. sampling from the data distribution is available, i.e., weighted sampling with $\mathbf{v} = \mathbf{w}$, our numerical results also indicate that CVaR fairness testing has a more favorable sample complexity than the baseline max-gap fairness testing.

VIII. Concluding Remarks

Conclusion. This paper proposes a multi-group fairness metric called conditional value-at-risk Fairness (CVaR fairness) inspired by conditional value-at-risk in finance. In fair machine learning, we are interested in testing if a given fairness metric is greater than a fixed threshold or equal to zero. We show that the usual multi-group fairness metrics (we call it max-gap fairness) have a high sample complexity, making testing for multi-group fairness infeasible. This motivates CVaR fairness, a metric that allows practitioners to relax fairness guarantees to decrease the sample complexity for reliably testing for multi-group fairness. We prove that it is possible to recover max-gap fairness from CVaR fairness by controlling one parameter in CVaR fairness. Next, we propose algorithms that test if the CVaR fairness is bigger than a threshold or equal to zero. We show that the proposed algorithms can reliably perform the hypothesis test with a more favorable sample complexity using an i.i.d. sample strategy (weighted sampling), improving the previous sample complexity in at least a square root factor. Moreover, we provided a non-i.i.d. sampling strategy (attribute specific sampling) that, paired with our proposed algorithm, can test for CVaR fairness with a test dataset size independent of the number of groups. Finally, we also provided a converse result on testing for CVaR fairness using weighted sampling, showing that the order of the sample complexity of the proposed algorithm is optimal under some assumptions. Our results show that by using CVaR fairness

and the proposed algorithms, ML practitioners can reliably test for multi-group fairness by trading off fairness guarantees for a more favorable sample complexity.

Ethical Considerations. We propose CVaR fairness as a relaxation of max-gap fairness. CVaR fairness achieves a more favorable sample complexity relative to max-gap fairness by providing a weaker fairness guarantee. We highlight that the parameter α captures this trade-off and can be tuned to fit both a fairness requirement and a sample size constraint. We recommend that practitioners choose this parameter responsibly and be mindful that small values of α may decrease fairness guarantees.

Future Work. We hope that future work explores the empirical implications of our testing proposal to demonstrate the trade-off between sample complexity and fairness guarantees from CVaR fairness. Additionally, by shifting from max-gap fairness to CVaR fairness, we are able to gradually improve sample complexity but this could come at the cost of possibly overlooking performance gaps for uniquely rare and unsampled groups. While we believe this should be rare in practice and our framework is tunable, it would be valuable for future work to empirically examine this on multiple real-world use cases. Moreover, our results only hold for binary group-specific losses, i.e., $L \in \{0, 1\}$. Hence, developing theoretical guarantees for multi-group fairness notions with non-binary L, such as multicalibration is another direction for future investigation. Finally, we point out that our results only hold for a finite (although as large as desired) number of groups. However, providing theoretical guarantees for testing for multi-group fairness in the presence of uncountable continuous groups (e.g., human height) is also of interest in the community [21], [22]. We highlight that CVaR fairness (Definition 3) can be straightforwardly generalized for continuous groups. However, the theoretical guarantees we provide don't automatically extend to handle such a definition. For this reason, we leave this contribution as future work.

IX. Acknowledgetments

The authors thank Flavien Prost (Google Research) and Alexandru Tifrea (ETHZ) for their helpful comments and discussions on an earlier version of this paper. We also thank the anonymous reviewers for their careful contribution.

References

[1] P. K. Roy, S. S. Chowdhary, and R. Bhatia, "A machine learning approach for automation of resume recommendation system," *Procedia Computer Science*, vol. 167, pp. 2318–2327, 2020.

[2] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in p2p lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and its Applications*, vol. 534, 2019.

[3] L. Brotcke, "Time to assess bias in machine learning models for credit decisions," *Journal of Risk and Financial Management*, vol. 15, 2022.

[4] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Learning to detect patterns of crime," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 515—-530.

[5] D. Csillag, L. Monteiro Paes, T. Ramos, J. Romano, R. Oliveira, R. Seixas, and P. Orenstein, "Amnioml: Amniotic fluid segmentation and volume prediction with uncertainty quantification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 15 494–15 502.

[6] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, 2021.

[7] "Civil rights act of 1964," in *Enrolled Acts and Resolutions of Congress*, General Records of the United States Government. U.S. Government Publishing Office, 1964.

[8] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, 2020.

[9] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Bias mitigation for machine learning classifiers: A comprehensive survey," *arXiv preprint arXiv:2207.07068*, 2022.

[10] A. Lowy, S. Baharlouei, R. Pavan, M. Razaviyayn, and A. Beirami, "A stochastic optimization framework for fair risk minimization," *Transactions on Machine Learning Research*, 2022, expert Certification. [Online]. Available: https://openreview.net/forum?id=P9Cj6RJmN2

[11] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. Winston Michalak, S. Asoodeh, and F. Calmon, "Beyond adult and COMPAS: Fair multi-class prediction via information projection," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=0e0es11XAIM

[12] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn, "Rényi fair inference," in *International Conference on Learning Representations*, 2019.

[13] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[14] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[15] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

[16] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[18] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 869–874.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 2012, pp. 214–226.

[20] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[21] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2564–2572.

[22] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (Computationally-identifiable) masses," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1939–1948.

[23] A. Wang, V. V. Ramaswamy, and O. Russakovsky, "Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation," *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[24] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 1918–1921.

[25] Y. Kong, "Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 485–494.

[26] L. Monteiro Paes, C. Long, B. Ustun, and F. Calmon, "On the epistemic limits of personalized prediction," in *Advances in Neural Information Processing Systems*, 2022, pp. 1979–1991.

[27] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[28] B. Hooks, *Ain't I a woman: Black women and feminism*. South End Press, 1992.

[29] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "An empirical study of rich subgroup fairness for machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 100–109. [Online]. Available: https://doi.org/10.1145/3287560.3287592

[30] E. Lett and W. G. La Cava, "Translating intersectionality to fair machine learning in health sciences," *Nature Machine Intelligence*, vol. 5, pp. 476—-479, 2023.

[31] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1929–1938.

[32] C. L. Canonne, "A survey on distribution testing: Your data is big. but is it blue?" *Theory of Computing*, pp. 1–100, 2020.

[33] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems 30*, 2017.

[34] P. Tchébychef, "Des valeurs moyennes (traduction du russe, n. de khanikof." *Journal de Mathématiques Pures et Appliquées*, pp. 177–184, 1867.

[35] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." *Journal für die reine und angewandte Mathematik*, pp. 210–271, 1909.

[36] R. Williamson and A. Menon, "Fairness risk measures," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6786–6797.

[37] A. Soen, I. M. Alabdulmohsin, S. Koyejo, Y. Mansour, N. Moorosi, R. Nock, K. Sun, and L. Xie, "Fair wrapping for black-box predictions," in *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

[38] S. Y. Meng and R. M. Gower, "A model-based method for minimizing CVaR and beyond," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 24 436–24 456. [Online]. Available: https://proceedings.mlr.press/v202/meng23a.html

[39] M. Falahatgar, A. Jafarpour, A. Orlitsky, V. Pichapati, and A. Theertha Suresh, "Faster algorithms for testing under conditional sampling," in *Proceedings of The 28th Conference on Learning Theory*, 2015, pp. 607–636.

[40] L. Le Cam, "Convergence of estimates under dimensionality restrictions," *The Annals of Statistics*, vol. 1, 1973.

[41] Y. Polyanskiy and Y. Wu, "Dualizing le cam's method for functional estimation, with applications to estimating the unseens," *arXiv preprint arXiv:1902.05616*, 2019.

[42] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: https://doi.org/10.21105/joss.03021

[43] Y. Ingster and I. Suslina, *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer, 01 2003, vol. 169.

APPENDIX A
TABLE OF NOTATION

TABLE I: Table of Notation

Symbol	Meaning
Δ^d	d-dimensional simplex
x	feature vector
y	true label
$\hat{y}$	predicted label
$\mathcal{G}$	set of all protected groups
g	specific protected group
z	tuple $z = (x, g, y, \hat{y})$
$\widetilde{P}_Z$	distribution of samples z
n	data budget
$\mathcal{D}_{\text{audit}}$	audit data $\mathcal{D}_{\text{audit}} = \{z_i\}_{i=0}^n$
$\mathbf{w} = (w_1, ..., w_{\lvert\mathcal{G}\rvert})$	stochastic vector of group distribution (also called group weight vector)
$\mathbf{v} = (v_1, ..., v_{\lvert\mathcal{G}\rvert})$	stochastic vector that represents group marginal from the audit dataset
ϵ	hypothesis test threshold
P_{error}	probability of error in the hypothesis test
L	quality of service function
$\overline{L}(\mathbf{w})$	average quality of service
$\Delta(g; P, L, \mathbf{w})$	per group fairness gap
`F_MaxGap`	max-gap fairness
α	group percentile threshold
$\texttt{F\_CVaR}_\alpha$	CVaR fairness
$\widehat{F}(\mathbf{w})$	CVaR fairness estimator

APPENDIX B
LOWER BOUNDS FOR HYPOTHESIS TESTING FOR MAX-GAP FAIRNESS

This section shows the lower bounds for the hypothesis test using the standard max-gap fairness metrics in Definition 2. Particularly, we focus on Equal Opportunity fairness. The results here were discussed in Section III.

Recall that the Hellinger divergence between two distributions P, Q on $\mathcal{Z}$ is given by

$$\mathrm{H}^2(P||Q) = \sum_{z\in\mathcal{Z}} \left(\sqrt{p(z)} - \sqrt{q(z)}\right)^2.$$

Lemma 2 (Close Distributions): For max-gap fairness metrics `F_MaxGap` (e.g., equal opportunity in Example 1 and statistical parity in Example 2) with quality of service function L and measuring the average quality of service using the distribution P such that $(L, P) \in \{(L_{EO}, P_{EO}), (L_{SP}, P_{SP})\}$. If $\epsilon \leq 0.5$ there exists a distribution $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1(\epsilon)$ such that:

$$\mathrm{H}^2(P_0||P_1) \leq 2 - 2\left(1 - \frac{2\epsilon^2}{|\mathcal{G}|}\right). \tag{59}$$

Proof: We will prove the desired result for equal opportunity described in Example 1. The result for other fairness metrics, such as statistical parity in Example 2, and equal odds [13], is analogous. We start by assuming that

$$P_0(Z|(\hat{Y}, Y, G)) = P_0(X|(\hat{Y}, Y, G)) = P_1(Z|(\hat{Y}, Y, G)) = P_1(X|(\hat{Y}, Y, G)),$$

i.e., the conditional distribution of X from P_0 and P_1 are equal

$$P_0(X|(\hat{Y}, Y, G)) = P_1(X|(\hat{Y}, Y, G)). \tag{60}$$

Now, from (60), we conclude that it is only necessary to bound the Hellinger divergence for the marginal distributions of $(\hat{Y}, Y, G)$, because

$$\mathrm{H}^2(P_0||P_1) = \mathbb{E}_Q\left[\left(\sqrt{\frac{P_1(Z)}{P_0(Z)}} - 1\right)^2\right] = \mathbb{E}\left[\left(\sqrt{\frac{P_1(\hat{Y}, Y, G)}{P_0(\hat{Y}, Y, G)}} - 1\right)^2\right].$$

Next, we compute the divergence of the marginals.

$$\mathrm{H}^2(P_0||P_1) \tag{61}$$

$$= \sum_{\substack{g\in\mathcal{G}\\ y\in\{0,1\}\\ \hat{y}\in\{0,1\}}} \left(\sqrt{P_0(\hat{Y}=\hat{y}, Y=y, G=g)} - \sqrt{P_1(\hat{Y}=\hat{y}, Y=y, G=g)}\right)^2 \tag{62}$$

$$= \sum_{\substack{g\in\mathcal{G}\\ y\in\{0,1\}\\ \hat{y}\in\{0,1\}}} \Bigg(\sqrt{P_0(\hat{Y}=\hat{y} \mid Y=y, G=g)P_0(Y=y \mid G=g)P_0(G=g)}$$

$$- \sqrt{P_1(\hat{Y}=\hat{y} \mid Y=y, G=g)P_1(Y=y \mid G=g)P_1(G=g)}\Bigg)^2 \tag{63}$$

$$= \sum_{\substack{g\in\mathcal{G}\\ \hat{y}\in\{0,1\}}} \Bigg(\sqrt{P_0(\hat{Y}=\hat{y} \mid Y=1, G=g)P_0(G=g)}$$

$$- \sqrt{P_1(\hat{Y}=\hat{y} \mid Y=1, G=g)P_1(G=g)}\Bigg)^2 \tag{64}$$

$$= \sum_{\substack{g\in\mathcal{G}\\ \hat{y}\in\{0,1\}}} P_0(\hat{Y}=\hat{y} \mid Y=1, G=g)w_g + P_1(\hat{Y}=\hat{y} \mid Y=1, G=g)w_g$$

$$- 2w_g\sqrt{P_1(\hat{Y}=\hat{y} \mid Y=1, G=g)P_0(\hat{Y}=\hat{y} \mid Y=1, G=g)} \tag{65}$$

$$= 2 - 2\sum_{g\in\mathcal{G}} w_g\Bigg(\sqrt{P_1(\hat{Y}=1 \mid Y=1, G=g)P_0(\hat{Y}=1 \mid Y=1, G=g)}$$

$$+ \sqrt{P_1(\hat{Y}=0 \mid Y=1, G=g)P_0(\hat{Y}=0 \mid Y=1, G=g)}\Bigg) \tag{66}$$

Now, take $w_g = \frac{1}{|\mathcal{G}|}$. Also, let P_0 and P_1 be :

$$P_0(\hat{Y}=\hat{y} \mid Y=1, G=g) = \frac{1}{2} \quad \forall g \in \mathcal{G}. \tag{67}$$

$$P_1(\hat{Y}=\hat{y} \mid Y=1, G=g) = \frac{1}{2} \quad \forall g \in \mathcal{G}\setminus\{g_1\}. \tag{68}$$

$$P_1(\hat{Y}=\hat{y} \mid Y=1, G=g_1) = \frac{1}{2} + \epsilon\frac{|\mathcal{G}|}{|\mathcal{G}|-1}. \tag{69}$$

Note that $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1(\epsilon)$.

Therefore, by plug in the distributions in (67) and (69) in the equality from (66), we have that

$$\begin{aligned}\mathrm{H}^2(P_0||P_1) &= 2-2\left(1+\frac{1}{|\mathcal{G}|}\left(\sqrt{\frac{1}{2}\left(\frac{1}{2}+\frac{|\mathcal{G}|}{|\mathcal{G}|-1}\epsilon\right)}+\sqrt{\frac{1}{2}\left(\frac{1}{2}-\frac{|\mathcal{G}|}{|\mathcal{G}|-1}\epsilon\right)}-1\right)\right)\\ &\leq 2-2\left(1+\frac{1}{|\mathcal{G}|}\left(\sqrt{\frac{1}{2}\left(\frac{1}{2}+\epsilon\right)}+\sqrt{\frac{1}{2}\left(\frac{1}{2}-\epsilon\right)}-1\right)\right)\\ &\leq 2-2\left(1-\frac{2\epsilon^2}{|\mathcal{G}|}\right),\end{aligned}$$

which completes the proof. □

Recall that the total variation divergence $\mathrm{TV}(.||.)$ takes a pair of distributions P and Q and returns

$$\mathrm{TV}(P||Q) \triangleq \frac{1}{2}\sum_{z\in\mathcal{Z}}|P(z)-Q(z)|. \tag{70}$$

Proposition 1: (Converse for Max-Gap Fairness). For max-gap fairness metrics `F_MaxGap` (e.g., equal opportunity in Example 1 and statistical parity in Example 2) with quality of service function L and measuring the average quality of service using the distribution P such that $(L,P)\in\{(L_{EO},P_{EO}),(L_{SP},P_{SP})\}$. If $\epsilon\in[0,0.5]$ and $w_g = P(G=g) = \frac{1}{|\mathcal{G}|}$ for all $g\in\mathcal{G}$, using n i.i.d. samples from P the minimum probability of error of performing an ϵ-test for `F_MaxGap` in Definition 1 is lower bounded by

$$2\inf_{\psi_\epsilon} P_{\text{error}} \geq 1-\left[2\left(1-\left(1-\frac{2\epsilon^2}{|\mathcal{G}|}\right)^n\right)\right]^{1/2}. \tag{23}$$

Furthermore, it is necessary to have access to n given in (24) i.i.d. samples from P to test if $\texttt{F\_MaxGap}=0$ or $\texttt{F\_MaxGap}\geq\epsilon$ correctly with probability 0.99.

$$n = \Omega\left(\frac{|\mathcal{G}|}{\epsilon^2}\right). \tag{24}$$

Proof: We will prove the desired result for equal opportunity described in Example 1. The result for other fairness metrics, such as statistical parity in Example 2, and equal odds [13], is analogous.

We start with n samples from two distributions $P_0\in\mathcal{P}_0$ and $P_1\in\mathcal{P}_1(\epsilon)$. Recall that the probability of error is given by

$$P_{\text{error}}(\psi) = \frac{1}{2}(P_0(\psi=1)+P_1(\psi=0)). \tag{71}$$

Therefore, by choosing the best decision function ψ, by the Le Cam lemma [40], and the fact that for all distributions P and Q

$$\mathrm{TV}(P||Q)\leq \mathrm{H}(P||Q),$$

we conclude that

$$\begin{aligned}\inf_{\psi} 2P_{\text{error}}(\psi) &= 1-\mathrm{TV}(P_0^n||P_1^n)\\ &\geq 1-\sqrt{\mathrm{H}^2(P_0^n||P_1^n)}\\ &\geq 1-\left(2\left(1-\left(1-\frac{1}{2}\mathrm{H}^2(P_0||P_1)\right)^n\right)\right)^{1/2}\end{aligned}$$

From Lemma 2 we have that there exists $P_0\in\mathcal{P}_0$ and $P_1\in\mathcal{P}_1(\epsilon)$ such that

$$\mathrm{H}^2(P_0||P_1)\leq 2-2\left(1-\frac{2\epsilon^2}{|\mathcal{G}|}\right), \tag{72}$$

and $\mathbf{w}$ uniform.

Taking the $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1(\epsilon)$ from Lemma 2, we have that

$$\begin{aligned}\inf_{\psi} 2P_{\text{error}}(\psi) &\geq 1 - \left(2\left(1 - \left(1 - \frac{1}{2}\mathrm{H}^2(P_0||P_1)\right)^n\right)\right)^{1/2} \\ &\geq 1 - \left(2\left(1 - \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|}\right)^n\right)\right)^{1/2} \end{aligned} \tag{73}$$

We now provide the lower bound on sample complexity. We start to differentiating between the distributions $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1(\epsilon)$ from Lemma 2. We want to ensure that

$$\delta \geq \inf_{\psi} P_e(\psi). \tag{74}$$

Next, we compute how large n needs to be to ensure that (73) holds.

$$2\delta \geq 1 - \left(2\left(1 - \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|}\right)^n\right)\right)^{1/2} \tag{75}$$

$$\Rightarrow 1 - \frac{(1-2\delta)^2}{2} \geq \left(1 - \frac{2\epsilon^2}{|\mathcal{G}|}\right)^n \tag{76}$$

$$\Rightarrow \ln\left(1 - \frac{(1-2\delta)^2}{2}\right) \geq n \ln\left(1 - \frac{2\epsilon^2}{|\mathcal{G}|}\right) \tag{77}$$

$$\Rightarrow \ln\left(1 - \frac{(1-2\delta)^2}{2}\right) \geq n \frac{-2\epsilon^2}{-2\epsilon^2 + |\mathcal{G}|} \tag{78}$$

$$\Rightarrow n \geq \ln\left(1 - \frac{(1-2\delta)^2}{2}\right) \frac{-2\epsilon^2 + |\mathcal{G}|}{-2\epsilon^2} \tag{79}$$

$$\Rightarrow n = \Omega\left(\frac{|\mathcal{G}|}{\epsilon^2}\right) \tag{80}$$

Hence, we have the desired result. □

Appendix C
Properties of CVaR Fairness

In this section, we prove the properties from CVaR fairness discussed in Section IV.

Lemma 3 (`F_MaxGap` is bounded): For all distributions P with support on $\mathcal{Z}$, if $L : \mathcal{Z} \to \{0, 1\}$ then

$$0 \leq \texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) \leq 1. \tag{81}$$

Proof: Recall that

$$\Delta(g) = \left|\mathbb{E}_P[L(Z)|G = g] - \overline{L}(\mathbf{w})\right| \tag{82}$$

$$= \left|\mathbb{E}_P[L(Z)|G = g] - \sum_{g \in \mathcal{G}} w_g \mathbb{E}_P[L(Z)|G = g]\right|. \tag{83}$$

Where $\overline{L}(\mathbf{w}) \in [0, 1]$ and $\mathbb{E}_P[L(Z)|G = g] \in [0, 1]$. Therefore,

$$\Delta(g) \leq 1 \Rightarrow \texttt{F\_MaxGap} = \max_g \Delta(g) \leq 1. \tag{84}$$

□

Proposition 3: ($\texttt{F\_CVaR}_\alpha$ is a lower bound for $\texttt{F\_MaxGap}$). For all $L : \mathcal{Z} \to \{0, 1\}$, distribution P with support on $\mathcal{Z}$ and distribution P with support on $\mathcal{Z}$, $\mathbf{w} \in \Delta^{|\mathcal{G}|}$, and $\alpha \in [0, 1)$ we have that

$$0 \leq \texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})) \leq \texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) \leq 1. \tag{33}$$

Proof: From the definition of CVaR fairness we know that:

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}) = \frac{1}{(1-\alpha)} \max_{\sum_{g \in Q \subset \mathcal{G}} w_g \leq (1-\alpha)} \sum_{g \in Q} w_g \Delta(g) \geq 0$$

because from the definition of $\Delta(g)$, $\Delta(g) \geq 0$ for all $g \in \mathcal{G}$ and $w_g \geq 0$ for all $g \in \mathcal{G}$.

Moreover, we have that:

$$\begin{aligned}
\texttt{F_CVaR}_\alpha(\mathbf{w}) &= \frac{1}{(1-\alpha)} \max_{\sum_{g \in Q \subset \mathcal{G}} w_g \leq (1-\alpha)} \sum_{g \in Q} w_g \Delta(g) \\
&\leq \frac{1}{(1-\alpha)} \max_{\sum_{g \in Q \subset \mathcal{G}} w_g \leq (1-\alpha)} \sum_{g \in Q} w_g \max_{g \in \mathcal{G}} \Delta(g) \\
&= \max_{g \in \mathcal{G}} \Delta(g) \max_{\sum_{g \in Q \subset \mathcal{G}} w_g \leq (1-\alpha)} \frac{1}{(1-\alpha)} \sum_{g \in Q} w_g \\
&\leq \max_{g \in \mathcal{G}} \Delta(g) \\
&= \texttt{F_MaxGap}(P, L, \overline{L}(\mathbf{w})).
\end{aligned}$$

□

Proposition 2: ($\texttt{F\_CVaR}_\alpha$ recovers $\texttt{F\_MaxGap}$). For all $\mathbf{w} \in \Delta^{|\mathcal{G}|}$, $L : \mathcal{Z} \to \{0, 1\}$, and any distribution P with support on $\mathcal{Z}$, there exists $\alpha^* \in [0, 1)$ such that

$$\texttt{F\_MaxGap}(P, L, \overline{L}(\mathbf{w})) = \texttt{F\_CVaR}_{\alpha^*}(\mathbf{w}; P, L, \overline{L}(\mathbf{w})). \tag{32}$$

When $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$ we have that $\alpha^* = 1 - \frac{1}{|\mathcal{G}|}$.

Proof: From proposition 3, we already know that

$$\texttt{F\_CVaR}_\alpha(\mathbf{w}) \leq \texttt{F\_MaxGap}(L, \overline{L}(\mathbf{w})).$$

However, by taking $\alpha = 1 - w_{g^*}$, the set $Q = \{w_{g^*}\}$ for some $g^* \in \arg\max_g \Delta(g)$ achieves the maximum

$$\begin{aligned}
\frac{1}{(1-\alpha)} \sum_{g \in Q} w_g \Delta(g) &= \frac{1}{w_{g^*}} w_{g^*} \Delta(g^*) \\
&= \Delta(g^*) \\
&= \texttt{F_MaxGap}(L, \overline{L}(\mathbf{w})).
\end{aligned}$$

Moreover, $\sum_{g \in Q \subset \mathcal{G}} w_g = w_{g^*} \leq w_{g^*} = 1 - \alpha$. Hence, the desired result follows from the fact that Q achieves the maximum because

$$\texttt{F\_MaxGap}(L, \overline{L}(\mathbf{w})) = \frac{1}{(1-\alpha)} \sum_{g \in Q} w_g \Delta(g) \tag{85}$$

$$\leq \frac{1}{(1-\alpha)} \max_{\sum_{g \in Q \subset \mathcal{G}} w_g \leq (1-\alpha)} \sum_{g \in Q} w_g \Delta(g) \tag{86}$$

$$= \texttt{F\_CVaR}_\alpha(\mathbf{w}) \leq \texttt{F\_MaxGap}(L, \overline{L}(\mathbf{w})). \tag{87}$$

Note that if $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$, then $\alpha = 1 - w_{g^*} = 1 - \frac{1}{|\mathcal{G}|}$.

□

APPENDIX D
UPPER BOUND FOR CVAR TESTING

This section proves the results in Section V-C.

Lemma 4: Under weighted sampling, if $nv_g \leq 1$ and $n \geq 2$,

$$\Pr(M_g \geq 1) \geq nv_g/e,$$

and

$$\Pr(M_g \geq 2) \geq n^2 v_g^2/4e.$$

Proof: Since $nv_g \leq 1$,

$$\Pr(M_s \geq 1) \geq \Pr(M_s = 1) = nv_g(1-v_g)^{n-1} \geq nv_g/e,$$

Similarly $nv_g \leq 1$,

$$\Pr(M_s \geq 2) \geq \Pr(M_s = 2) = n(n-1)/2(v_g)^2(1-v_g)^{n-2} \geq n^2 v_s^2/4e,$$

where the last inequality follows from $n \geq 2$. □

Lemma 5 (Computing Average of First Order Estimation): Define $\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g} = 0$ if $M_g = 0$ and recall that $z = (x, g, y)$. If $L : \mathcal{Z} \to \{0, 1\}$, then

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right] = \mathbb{E}[L(z) \mid G = g] \Pr(M_g \geq 1) \tag{88}$$

Proof: By the conditional law of expectation,

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right]\right].$$

If $m_g = 0$, by definition,

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right] = 0. \tag{89}$$

If $m_g \geq 1$, then

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right] = \sum_{i=1}^{m_g} \frac{\mathbb{E}\left[L(z_i)|\, M_g = m_g\right]}{m_g} \tag{90}$$

$$= \sum_{i=1}^{m_g} \frac{\mathbb{E}\left[L(z)|\, G = g\right]}{m_g} \tag{91}$$

$$= \mathbb{E}\left[L(z)|\, G = g\right] \tag{92}$$

Combining the above two equations yields

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right] = \mathbb{E}\left[L(z)|\, G = g\right] \mathbb{I}[M_g \geq 1].$$

Taking the expectation on both sides w.r.t. M_g yields the lemma. □

Lemma 6 (Computing Variance of First Order Estimation): Define $\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g} = 0$ for $M_g = 0$ and recall that $z = (x, g, y)$. If $L : \mathcal{Z} \to \{0, 1\}$, then

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right] \leq 2 \Pr(M_g \geq 1)$$

Proof: By the law of total variance,

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right] = \mathbb{E}\left[\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g\right]\right] + \text{Var}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g\right]\right]. \tag{93}$$

We first compute the second term in the RHS of the above equation (93). By Lemma 5 we have that:

$$\begin{aligned}\text{Var}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g\right]\right] &= \text{Var}\left[\mathbb{E}\left[L(z)|\, G=g\right]\mathbb{I}[M_g \geq 1]\right] &(94)\\ &= \mathbb{E}\left[L(z)|\, G=g\right]^2 \text{Var}\left[\mathbb{I}[M_g \geq 1]\right] &(95)\\ &\leq \mathbb{E}\left[L(z)|\, G=g\right]^2 \Pr\left(M_g \geq 1\right) &(96)\\ &\leq \Pr\left(M_g \geq 1\right) &(97)\end{aligned}$$

Now, let's bound the first term in the first term in the RHS of (93). Let's start by bounding $\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right]$.

If $m_g = 0$ then:

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g\right] = 0$$

If $m_g \geq 1$, then by Popoviciu's inequality on variances:

$$\begin{aligned}\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g\right] &\leq \frac{1}{4}\left(\max_{z_i}\left(\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right) - \min_{z_i}\left(\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right)\right)^2\\ &\leq \frac{1}{4}(0-(-1)1)^2 = \frac{1}{4},\end{aligned}$$

where the maximum is achieved when all z_i are such that $L(z_i) = 1$ and the minimum is achieved when $L(z_i) = 0$ — note that the bound holds when $M_g \neq 0$.

By the two last equations, we conclude that:

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right] \leq \frac{1}{4}\mathbb{I}[m_g \geq 1] \tag{98}$$

and by taking expectations on both sides:

$$\mathbb{E}\left[\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\middle| M_g = m_g\right]\right] \leq \frac{1}{4}\Pr(m_g \geq 1) \tag{99}$$

Combining equation 93, 97, and 99 we conclude that:

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\right] \leq \left(1+\frac{1}{4}\right)\Pr(m_g \geq 1) \leq 2\Pr(m_g \geq 1) \tag{100}$$

□

Lemma 7 (Computing Average of Second Order Estimation): Define $\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1} = 0$ if $M_g = 0, 1$, and recall that $z = (x, g, y)$. If $L : \mathcal{Z} \to \{0, 1\}$, then

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right] = \mathbb{E}[L(z) \mid G=g]^2 \Pr(M_g \geq 2)$$

Proof: By the conditional law of expectation,

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right] =\mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g=m_g\right]\right] \tag{101}$$

$$\tag{102}$$

If $m_g \in \{0,1\}$ then:

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g=m_g\right] = 0 \tag{103}$$

If $m_g \geq 2$ then:

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g=m_g\right] \tag{104}$$

$$=\mathbb{E}\left[\frac{\sum_{i=1}^{m_g} L(z_i)}{m_g}\frac{\sum_{i=1}^{m_g} L(z_i)-1}{m_g-1}\middle| M_g=m_g\right] \tag{105}$$

$$=\mathbb{E}\left[\frac{\left(\sum_{i=1}^{m_g} L(z_i)\right)^2-\sum_{i=1}^{m_g} L(z_i)}{m_g(m_g-1)}\middle| M_g=m_g\right] \tag{106}$$

$$=\mathbb{E}\left[\frac{\sum_{i=1}^{m_g} L(z_i)^2+\sum_{i\neq j} L(z_i)L(z_j)-\sum_{i=1}^{m_g} L(z_i)}{m_g(m_g-1)}\middle| M_g=m_g\right] \tag{107}$$

$$=\frac{m_g\mathbb{E}\left[L(z)^2-L(z)|G=g\right]+m_g(m_g-1)\mathbb{E}\left[L(z)|G=g\right]^2}{m_g(m_g-1)} \tag{108}$$

$$=\mathbb{E}\left[L(z)|G=g\right]^2, \tag{109}$$

where the last inequality comes from the fact that $\mathbb{E}\left[L(z)^2-L(z)|G=g\right]=0$ because $L(z)\in\{0,1\}$ therefore $L(z)^2-L(z)=0$ for all z.

Combining (103) and (109) yields

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g=m_g\right] =\mathbb{E}\left[L(z)|G=g\right]^2\mathbb{I}[M_g\geq 2]. \tag{110}$$

Taking the expectation on both sides w.r.t. M_g give us the lemma. □

Lemma 8 (Computing Variance of Second Order Estimation): Define $\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}=0$ if $M_g=0,1$, and recall that $z=(x,g,y)$. If $L:\mathcal{Z}\to\{0,1\}$, then

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right] \leq 2\Pr(M_g\geq 2)$$

Proof: By the law of total variance,

$$\begin{aligned}\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right] &= \mathbb{E}\left[\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g\right]\right] \\ &+\text{Var}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g\right]\right].\end{aligned} \tag{111}$$

We first compute the second term in the RHS of the above equation (111). By Lemma 7 we have that:

$$\begin{aligned}
&\text{Var}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g\right]\right] \\
&= \text{Var}\left[\left(\frac{\mathbb{E}\left[L(z)^2 - L(z)|G=g\right]}{(m_g-1)} + \mathbb{E}\left[L(z)|G=g\right]^2\right)\mathbb{I}[m_g \geq 2]\right] \qquad (112)
\end{aligned}$$

Note that $\mathbb{E}\left[L(z)^2 - L(z)|G=g\right] = 0$.
Now, we can bound Eq. 112 by

$$\begin{aligned}
&= \text{Var}\left[\left(\frac{\mathbb{E}\left[L(z)^2 - L(z)|G=g\right]}{(m_g-1)} + \mathbb{E}\left[L(z)|G=g\right]^2\right)\mathbb{I}[m_g \geq 2]\right] \\
&= \text{Var}\left[\left(\mathbb{E}\left[L(z)|G=g\right]^2\right)\mathbb{I}[m_g \geq 2]\right] \qquad (113) \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[L(z)|G=g\right]^2\right)^2\mathbb{I}[m_g \geq 2]\right] - \mathbb{E}\left[\left(\mathbb{E}\left[L(z)|G=g\right]^2\right)\mathbb{I}[m_g \geq 2]\right]^2 \qquad (114) \\
&\leq \mathbb{E}\left[\left(\mathbb{E}\left[L(z)|G=g\right]^2\right)^2\mathbb{I}[m_g \geq 2]\right] \qquad (115) \\
&\leq \left|\left|\left(\mathbb{E}\left[L(z)|G=g\right]^2\right)^2\right|\right|_\infty \mathbb{E}\left[\mathbb{I}[m_g \geq 2]\right] \\
&\leq \Pr(m_g \geq 2) \qquad (116)
\end{aligned}$$

Where the last inequality comes from the fact that $L \in \{0,1\}$.

Now, let's bound the first term in the first term in the RHS of (111). Let's start by bounding $\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g = m_g\right]$.

If $m_g \in \{0,1\}$ then:

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g = m_g\right] = 0$$

If $m_g \geq 2$ then, by Popoviciu's inequality on variances:

$$\begin{aligned}
\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g = m_g\right] &\leq \frac{1}{4}\left(\max_{z_i}\left(\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right)\right. \\
&\quad \left. - \min_{z_i}\left(\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right)\right)^2 \\
&\leq \frac{1}{4}(1-0)^2 = \frac{1}{4}
\end{aligned}$$

By the two last equations, we conclude that:

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g = m_g\right] \leq \frac{1}{4}\mathbb{I}[m_g \geq 2] \qquad (117)$$

By taking averages on both sides, we get:

$$\mathbb{E}\left[\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\middle| M_g\right]\right] \leq \frac{1}{4}\Pr(M_g \geq 2). \qquad (118)$$

Bounding the RHS of (111) by using the result in (116) and (118) we conclude that

$$\text{Var}\left[\frac{\sum_{i=1}^{M_g} L(z_i)}{M_g}\frac{\sum_{i=1}^{M_g} L(z_i)-1}{M_g-1}\right] \leq \frac{5}{4}\Pr(M_g \geq 2) \tag{119}$$

$$\leq 2\Pr(M_g \geq 2). \tag{120}$$

□

Lemma 9: Denote $E_g = \mathbb{E}\left[L(z)|G=g\right]$, if $L : \mathcal{Z} \to \{0,1\}$, then

$$\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2 \geq (1-\alpha)(\texttt{F\_CVaR}_\alpha(w))^2. \tag{121}$$

Proof: The proof follows from algebraic manipulation as follows.

$$(1-\alpha)\left(\texttt{F\_CVaR}_\alpha(w)\right)^2 + \overline{L}(\mathbf{w})^2 = (1-\alpha)\left(\max_{\sum w_g \leq 1-\alpha} \sum \frac{w_g}{(1-\alpha)}\Delta(g)\right)^2 + \overline{L}(\mathbf{w})^2 \tag{122}$$

$$\leq (1-\alpha)\left(\sum_{g\in Q_\alpha} \frac{\sqrt{w_g}}{\sqrt{(1-\alpha)}}\frac{\sqrt{w_g}\Delta(g)}{\sqrt{(1-\alpha)}}\right)^2 + \overline{L}(\mathbf{w})^2$$

$$\leq (1-\alpha)\left(\sum_{g\in Q_\alpha} \frac{w_g}{(1-\alpha)}\right)\left(\sum_{g\in Q_\alpha} \frac{w_g\Delta(g)^2}{(1-\alpha)}\right) + \overline{L}(\mathbf{w})^2 \tag{123}$$

$$\leq \sum_{g\in Q_\alpha} w_g\Delta(g)^2 + \overline{L}(\mathbf{w})^2$$

$$= \sum_{g\in Q_\alpha} w_g(E_g - \overline{L}(\mathbf{w}))^2 + \overline{L}(\mathbf{w})^2$$

$$\leq \sum_{g\in\mathcal{G}} w_g(E_g - \overline{L}(\mathbf{w}))^2 + \overline{L}(\mathbf{w})^2$$

$$= \sum_{g\in\mathcal{G}} w_g E_g^2 - 2\overline{L}(\mathbf{w})\left(\sum_{g\in\mathcal{G}} w_g E_g - \overline{L}(\mathbf{w})\right)$$

$$= \sum_{g\in\mathcal{G}} w_g E_g^2,$$

where the inequality in 123 follows by invoking Cauchy-Schwartz inequality. □

Lemma 10: If $L : \mathcal{Z} \to \{0,1\}$ is the quality of service function and $\overline{L}(\mathbf{w})$ be the average quality of service and $w_g \leq 1-\alpha$ for all $g \in \mathcal{G}$, then

(a) If $\texttt{F\_CVaR}_\alpha(w) = 0$, then

$$\Pr\left(\widehat{F} > (1-\alpha)\epsilon^2/2\right) \leq \sum_g \frac{64w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]}.$$

Similarly,

(b) if $\texttt{F\_CVaR}_\alpha(w) \geq \epsilon$, then

$$\Pr\left(\widehat{F} < (1-\alpha)\epsilon^2/2\right) \leq \sum_g \frac{64w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]}.$$

Proof:

Denote $E_g = \mathbb{E}\left[L(z)|G=g\right]$.

Recall that the estimator is given by

$$\widehat{F} = \sum_{g\in\mathcal{G}} \frac{w_g}{P\left[M_s \geq 2\right]} \frac{\sum_{i=1}^{M_g} L(z_i)}{M_g} \frac{(\sum_{i=1}^{M_g} L(z_i) - 1)}{M_g - 1} - \left(\sum_{g\in\mathcal{G}} \frac{w_g}{P\left[M_s \geq 1\right]} \frac{\sum_{j=1}^{M_g} L(z_j)}{M_g}\right)^2 \tag{124}$$

We first show the result under the null hypothesis, i.e., when $\texttt{F\_CVaR}_\alpha(w) = 0$. Note that this condition implies that $\left(\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2\right) = 0$, because as all $w_g \leq 1-\alpha$ we have that if $E_g \neq \overline{L}(\mathbf{w})$ for some $g \in \mathcal{G}$, then $\Delta(g) > 0$ and $\texttt{F\_CVaR}_\alpha(w) \geq \frac{w_g \Delta(g)}{1-\alpha} > 0$.

$$\Pr(\widehat{F} > (1-\alpha)\epsilon^2/2) \tag{125}$$

$$= \Pr\left(\widehat{F} - \left(\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2\right) > (1-\alpha)\epsilon^2/2\right) \tag{126}$$

$$\leq \Pr\left(\sum_{g\in\mathcal{G}} \frac{w_g \sum_{i=1}^{M_g} L(z_i)(\sum_{i=1}^{M_g} L(z_i) - 1)}{P\left[M_s \geq 2\right] M_g(M_g-1)} - \sum_{g\in\mathcal{G}} w_g E_g^2 > \frac{(1-\alpha)\epsilon^2}{4}\right)$$

$$+ \Pr\left(\overline{L}(\mathbf{w})^2 - \left(\sum_{g\in\mathcal{G}} \frac{w_g}{P\left[M_s \geq 1\right]} \frac{\sum_{j=1}^{M_g} L(z_j)}{M_g}\right)^2 > \frac{(1-\alpha)\epsilon^2}{4}\right) \tag{127}$$

$$\leq \Pr\left(\sum_{g\in\mathcal{G}} \frac{w_g \sum_{i=1}^{M_g} L(z_i)(\sum_{i=1}^{M_g} L(z_i) - 1)}{P\left[M_s \geq 2\right] M_g(M_g-1)} - \sum_{g\in\mathcal{G}} w_g E_g^2 > \frac{(1-\alpha)\epsilon^2}{4}\right)$$

$$+ \Pr\left(\left(\sum_{g\in\mathcal{G}} \frac{w_g}{P\left[M_s \geq 1\right]} \frac{\sum_{j=1}^{M_g} L(z_j)}{M_g}\right) < \overline{L}(\mathbf{w}) - \frac{(1-\alpha)\epsilon^2}{8}\right), \tag{128}$$

where the last inequality follows by observing that $a^2 - b^2 > c$ implies that $a^2 - c > b^2$ which implies that $\sqrt{a^2-c} > b$ because b is positive. Then, we conclude by showing that

$$\sqrt{a^2-c} \leq a - \frac{c}{2} \tag{129}$$

$$\iff a^2 - c \leq \left(a - \frac{c}{2}\right)^2 \tag{130}$$

$$\iff a^2 - c \leq a^2 - ac + \frac{c^4}{4} \tag{131}$$

$$\iff -c \leq -ac + \frac{c^4}{4} \tag{132}$$

$$\iff c(a-1) \leq \frac{c^4}{4}, \tag{133}$$

the last inequality holds because $a \leq 1$ and $c > 0$.

Next, we bound each of the terms on the right-hand side of the above equation using Chebyshev's inequality.

First, define $\widehat{E}_1$ such that:

$$\widehat{E}_1 = \sum_{g\in\mathcal{G}} \frac{w_g \sum_{i=1}^{M_g} L(z_i)(\sum_{i=1}^{M_g} L(z_i) - 1)}{P\left[M_g \geq 2\right] M_g(M_g-1)} - \sum_{g\in\mathcal{G}} w_g E_g^2 \tag{134}$$

By Lemma 7,

$$\mathbb{E}[\widehat{E}_1] = 0,$$

and by Lemma 8

$$\text{Var}(\widehat{E}_1) \leq \sum_g \frac{2w_g^2}{P\left[M_g \geq 2\right]}.$$

Hence by Chebyshev's inequality,

$$\Pr\left(\widehat{E}_1 > (1-\alpha)\epsilon^2/4\right) \leq \Pr\left(\widehat{E}_1 - \mathbb{E}[\widehat{E}_1] > (1-\alpha)\epsilon^2/4\right) \tag{135}$$

$$\leq \Pr\left(\left|\widehat{E}_1 - \mathbb{E}[\widehat{E}_1]\right| > (1-\alpha)\epsilon^2/4\right) \tag{136}$$

$$\leq \sum_g \frac{32w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} \tag{137}$$

Additionally, denote $\widehat{E}_2$ by:

$$\widehat{E}_2 = \sum_{g\in\mathcal{G}} \frac{w_g}{P\left[M_g \geq 1\right]} \frac{\sum_{j=1}^{M_g} L(z_j)}{M_g} - \overline{L}(\mathbf{w}) \tag{138}$$

Similar to the previous argument, we conclude that:

$$\mathbb{E}[\widehat{E}_2] = 0,$$

and by Lemma 6

$$\text{Var}(\widehat{E}_1) \leq \sum_g \frac{2w_g^2}{P\left[M_g \geq 1\right]}.$$

Hence, by applying Chebyshev's inequality,

$$\begin{aligned}\Pr\left(\widehat{E}_2 < -(1-\alpha)\epsilon^2/8\right) &= \Pr\left(\mathbb{E}[\widehat{E}_2] - \widehat{E}_2 > (1-\alpha)\epsilon^2/8\right)\\ &\leq \Pr\left(\left|\mathbb{E}[\widehat{E}_2] - \widehat{E}_2\right| > (1-\alpha)\epsilon^2/8\right)\\ &\leq \frac{128}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]}\end{aligned}$$

Hence,

$$\Pr(\widehat{F} > (1-\alpha)\epsilon^2/2) \leq \sum_g \frac{32w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]},$$

completing the proof.

Next, we show the result under the hypothesis H_1, i.e., when $\texttt{F\_CVaR}_\alpha \geq \epsilon$. We are interested in bounding:

$$\Pr(\widehat{F} < (1-\alpha)\epsilon^2/2) \tag{139}$$

From Lemma 9, we have that

$$\begin{aligned}\Pr(\widehat{F} < (1-\alpha)\epsilon^2/2) &\leq \Pr\left(\widehat{F} - \left(\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2\right) < -(1-\alpha)\epsilon^2 + (1-\alpha)\epsilon^2/2\right)\\ &\leq \Pr\left(\widehat{F} - \left(\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2\right) < -(1-\alpha)\epsilon^2/2\right)\\ &\leq \Pr\left(-\widehat{F} + \left(\sum_{g\in\mathcal{G}} w_g E_g^2 - \overline{L}(\mathbf{w})^2\right) > (1-\alpha)\epsilon^2/2\right) \qquad (140)\end{aligned}$$

Using the bound in (140), and the same procedure we used in for bounding the probability of error when $\texttt{F\_CVaR}_\alpha = 0$, we can show that under $\texttt{F\_CVaR}_\alpha \geq \epsilon$

$$\Pr(\widehat{F} > (1-\alpha)\epsilon^2/2) \leq \sum_g \frac{32 w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4} \sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]},$$

□

Theorem 1: (Achievability with weighted sampling). Let $\mathbf{w} \in \Delta^{|\mathcal{G}|}$ be the group weight vector such that $\max_g w_g \leq 1-\alpha$, $L : \mathcal{Z} \to \{0,1\}$ be the quality of service function, and $\overline{L}(\mathbf{w})$ is the average quality of service. Suppose we have access to n i.i.d. samples using, i.e., we used weighted sampling in Algorithm 1. Then,

$$Q_g(k) = \Pr(\text{Bin}(n, v_g) = k),$$

where $\mathbf{v} = (v_1, ..., v_{|\mathcal{G}|})$ is the group marginal from P_Z that we sampled i.i.d. to generate $\mathcal{D}_{\text{audit}}$.

The probability of error (P_{error} in (5)) for Algorithm 1 to differentiate $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ from $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ is such that

$$P_{\text{error}} = O\left(\frac{1}{(1-\alpha)^2\epsilon^4 n^2} \sum_g \frac{w_g^2}{v_g^2} + \frac{1}{n(1-\alpha)^2\epsilon^4} \sum_g \frac{w_g^2}{v_g}\right). \tag{37}$$

Proof: Using Algorithm 1, we have that:

$$P_{\text{error}} = \Pr\left(\widehat{F} \geq \frac{\alpha\epsilon^2}{2} \middle| \texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0\right) \Pr\left(\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0\right) \tag{141}$$

$$+ \Pr\left(\widehat{F} < \frac{\alpha\epsilon^2}{2} \middle| \texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon\right) \Pr\left(\texttt{F\_CVaR}_\alpha(\mathbf{w}) \geq \epsilon\right) \tag{142}$$

then, by Lemma 10, we can bound the RHS (142) and conclude that

$$P_{\text{error}} \leq \sum_g \frac{32 w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4} \sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]} \tag{143}$$

By using weighted sampling we can lower bound the probabilities $P\left[M_g \geq 2\right]$ and $P\left[M_g \geq 1\right]$ as in Lemma 4. Using this lower bound, we can upper bound the RHS of (143) and conclude that

$$P_{\text{error}} \leq \sum_g \frac{128 e w_g^2}{(1-\alpha)^2\epsilon^4 n^2 (v_g)^2} + \frac{128 e}{(1-\alpha)^2\epsilon^4} \sum_g \frac{w_g^2}{n v_g} \tag{144}$$

Hence, we have the desired result

$$P_{\text{error}} = O\left(\sum_g \frac{w_g^2}{(1-\alpha)^2\epsilon^4 n^2 (v_g)^2} + \frac{1}{(1-\alpha)^2\epsilon^4} \sum_g \frac{w_g^2}{n v_g}\right). \tag{145}$$

□

Corollary 1: (Sample complexity of Algorithm 1). Under the same assumption of Theorem 1 (using weighted sampling). With probability at least 0.99 [4], comparing $\widehat{F}(\mathbf{w})$ to $(1-\alpha)\epsilon^2/2$, one can differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ using at most the following number of samples

$$n = O\left(\frac{\left(\sum_{g\in\mathcal{G}} \frac{w_g^2}{v_g^2}\right)^{1/2}}{\epsilon^2(1-\alpha)} + \frac{\sum_g \frac{w_g^2}{v_g}}{\epsilon^4(1-\alpha)^2}\right). \tag{39}$$

[4]Because the probability of success in the hypothesis test is not the main focus of our theoretical analysis, we leave the dependence on it to the appendix and only show the results in the main paper for a probability of error of 0.01.

Moreover, when the model auditor can control the choice of group marginal in the test dataset and set $v_g = \frac{w_g^{2/3}}{\sum_{g^*} w_{g^*}^{2/3}}$, it is only necessary to have access to n samples such that

$$n = O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\epsilon^2(1-\alpha)} + \frac{2^{\frac{1}{3}H_{2/3}(w)}}{\epsilon^4(1-\alpha)^2}\right). \tag{40}$$

Proof: The bound on n for general $\mathbf{v}$ and $\mathbf{w}$ follows from the error probability in Theorem 1. We now prove the result when $v_g \propto w_g^{2/3}$. From Theorem 1 we have that

$$P_{\text{error}} \leq \sum_g \frac{128 e w_g^2}{(1-\alpha)^2\epsilon^4 n^2 (v_g)^2} + \frac{128e}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{n v_g} \tag{146}$$

$$= 128e\frac{\left(\sum_g w_g^{2/3}\right)^3}{(1-\alpha)^2\epsilon^4 n^2} + \frac{128e}{(1-\alpha)^2\epsilon^4}\frac{(\sum_g w_g^{2/3})\sum_g w_g^{4/3}}{n} \tag{147}$$

$$\leq 128e\frac{\left(\sum_g w_g^{2/3}\right)^3}{(1-\alpha)^2\epsilon^4 n^2} + \frac{128e}{(1-\alpha)^2\epsilon^4}\frac{(\sum_g w_g^{2/3})}{n} \tag{148}$$

only need to ensure that

$$\delta \geq 128e\frac{\left(\sum_g w_g^{2/3}\right)^3}{(1-\alpha)^2\epsilon^4 n^2} + \frac{128e}{(1-\alpha)^2\epsilon^4}\frac{(\sum_g w_g^{2/3})}{n} \tag{149}$$

Therefore, it is only necessary to use n samples to test for CvaR fairness reliably, where n is such that

$$n = O\left(\frac{\left(\sum_g w_g^{2/3}\right)^{3/2}}{\delta\epsilon^2(1-\alpha)} + \frac{\left(\sum_g w_g^{2/3}\right)}{\delta\epsilon^4(1-\alpha)^2}\right) \tag{150}$$

$$\iff n = O\left(\frac{2^{\frac{3}{2}\log(\sum_g w_g^{2/3})}}{\delta\epsilon^2(1-\alpha)} + \frac{2^{\log(\sum_g w_g^{2/3})}}{\delta\epsilon^4(1-\alpha)^2}\right) \tag{151}$$

$$\iff n = O\left(\frac{2^{\frac{1}{2}H_{2/3}(w)}}{\delta\epsilon^2(1-\alpha)} + \frac{2^{\frac{1}{3}H_{2/3}(w)}}{\delta\epsilon^4(1-\alpha)^2}\right) \tag{152}$$

Setting $\delta = 0.01$ yields the result.

□

Theorem 2: (Achievability with attribute specific sampling). Let $\mathbf{w}$ be such that $\max_g w_g \leq c \cdot \epsilon^4$ for a sufficiently small constant c and $\max_g w_g \leq 1-\alpha$, $L : \mathcal{Z} \to \{0,1\}$ the quality of service function, and $\overline{L}(\mathbf{w})$ the average quality of service. Using attribute specific sampling in Algorithm 1, i.e.

$$Q_g(k) = \Pr\left(\text{Ber}(\min(1, \gamma w_g))\frac{n}{\gamma} = k\right),$$

where $\gamma = \frac{n}{2}$.

The probability of error (P_{error} in (5)) for Algorithm 1 to differentiate $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ from $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ is bounded by:

$$P_{\text{error}} = O\left(\frac{1}{\epsilon^4(1-\alpha)^2 n}\right). \tag{42}$$

Proof: Let $\delta = 0.01$. As in the proof of Theorem 1, we have that

$$P_{\text{error}} \leq \sum_g \frac{128 w_g^2}{(1-\alpha)^2\epsilon^4 P\left[M_g \geq 2\right]} + \frac{128}{(1-\alpha)^2\epsilon^4}\sum_g \frac{w_g^2}{P\left[M_g \geq 1\right]} \tag{153}$$

Using attribute specific sampling with $\gamma = \frac{n}{2}$ we conclude that $P[M_g \geq 1] = P[M_g \geq 2] = \min(1, \gamma w_g)$. Therefore, we can bound the probability of error in (153) as:

$$P_{\text{error}} \leq \sum_g \frac{128 w_g^2}{(1-\alpha)^2 \epsilon^4 \min(1, \gamma w_g)} + \frac{128}{(1-\alpha)^2 \epsilon^4} \sum_g \frac{w_g^2}{\min(1, \gamma w_g)} \tag{154}$$

$$= \sum_g \frac{256 w_g}{(1-\alpha)^2 \epsilon^4 n} + \frac{256}{(1-\alpha)^2 \epsilon^4} \sum_g \frac{w_g}{n} \tag{155}$$

$$\leq \frac{256}{(1-\alpha)^2 \epsilon^4 n} \tag{156}$$

Therefore, we conclude that

$$P_{\text{error}} = O\left(\frac{1}{\epsilon^4 (1-\alpha)^2 n}\right). \tag{157}$$

□

Corollary 2: (Sample complexity of Algorithm 1 using attribute specific sampling). Let $\mathbf{w}$ be such that $\max_g w_g \leq c \cdot \epsilon^4$ for a sufficiently small constant c and $\max_g w_g \leq 1 - \alpha$. With probability at least 0.99, comparing $\widehat{F}(\mathbf{w})$ to $(1-\alpha)\epsilon^2/2$, one can differentiate between $\texttt{F\_CVaR}_\alpha(\mathbf{w}) = 0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w}) > \epsilon$ using at most the following number of samples n:

$$n = O\left(\frac{1}{\epsilon^4 (1-\alpha)^2}\right).$$

Proof: Let $\delta = 0.01$. From Theorem 2, we just need to ensure that:

$$P_{\text{error}} \leq \frac{256}{(1-\alpha)^2 \epsilon^4 n} \leq \delta \tag{158}$$

Therefore,

$$\frac{256}{(1-\alpha)^2 \epsilon^4 \delta} \leq n \tag{159}$$

Concluding the proof. □

Appendix E
Converse of Testing for CVaR under Weighted Sampling

Lemma 1: [Close Mixture of Distributions] For CVaR fairness metrics $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ with quality of service function L and measuring the average quality of service using the distribution P such that $(L, P) \in \{(L_{EO}, P_{EO}), (L_{SP}, P_{SP})\}$. Consider the hypothesis regions $\mathcal{P}_0$ (3) and $\mathcal{P}_1(\epsilon)$ (4). Assume that $\mathbf{w}$ is a uniform distribution over all groups, i.e., $\mathbf{w} = \left(\frac{1}{|\mathcal{G}|}, ..., \frac{1}{|\mathcal{G}|}\right)$, For all $\alpha \in [0, 1)$ and $\frac{\alpha |\mathcal{G}|^{1/3}}{4} \geq \epsilon > 0$ there exist a sequence of distributions $P_0 \in \mathcal{P}_0$ and $\{P_u\}_{u \in \mathcal{U}} \subset \mathcal{P}_1(\epsilon)$ indexed by elements of the set $\mathcal{U}$ such that if $u \sim \pi$ we have that

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right] \middle|\middle| P_0^n\right) \leq e^{\frac{128(1-\alpha) n^2 \epsilon^4}{\alpha^4 |\mathcal{G}|}} - 1. \tag{44}$$

Proof: We will prove the desired result for equal opportunity described in Example 1. The result for other fairness metrics, such as statistical parity in Example 2, and equal odds [13], is analogous.

Let $\mathbf{w}$ be the uniform distribution over all groups. Note that under this definition $\mathbf{v} = \mathbf{w}$ for any η. Let $Q \subset \mathcal{G}$ be such that $|Q| = \lfloor (1-\alpha)|\mathcal{G}| \rfloor$. We define the set $\mathcal{U}(Q)$ to be

$$\mathcal{U}(Q) \triangleq \{-1, 1\}^{|Q|}. \tag{160}$$

$\mathcal{U}(Q)$ is the set of vector with entries equal to $+1$ or -1. We will use $\mathcal{U}(Q)$ to perturb the distributions of groups $g \in Q$. Denote the group distribution vector $\mathbf{w} = (w_1, ..., w_{|\mathcal{G}|})$. And define

$$P_0(G = g) = w_g, \tag{161}$$
$$P_u(G = g) = w_g \quad \forall u \in \mathcal{U}(Q). \tag{162}$$

Hence, we define the conditional distribution of $\hat{Y}$ given G accordingly with P_0 and P_u to be

$$P_0(\hat{Y} = 1|G = g, Y = 0) = \frac{1}{2} \qquad \forall g \in \mathcal{G} \tag{163}$$
$$P_u(\hat{Y} = 1|G = g, Y = 0) = \frac{1}{2} \qquad \forall g \notin Q \tag{164}$$
$$P_u(\hat{Y} = 1|G = g, Y = 0) = \frac{1}{2} + \frac{\tau \epsilon_g u_g}{w_g} \qquad \forall g \in Q. \tag{165}$$

Where $u = (u_1, ..., u_{|Q|})$, ϵ_g is such that

$$\epsilon_g \triangleq \frac{\epsilon w_g^{2/3}}{\left(\sum_{g \in Q} w_g^{2/3}\right)}, \tag{166}$$

and τ is given by

$$\tau \triangleq \frac{1-\alpha}{\alpha}. \tag{167}$$

Notice that we need to show that the probability distributions are well defined. To do so, it is only necessary to ensure that $\frac{\tau \epsilon_g}{w_g} \leq \frac{1}{2}$.

$$\tau \frac{\epsilon_g}{w_g} = \tau \frac{\epsilon w_g^{-1/3}}{\left(\sum_{g' \in Q} w_{g'}^{2/3}\right)} = \tau \frac{\epsilon |\mathcal{G}|^{1/3}}{\lfloor (1-\alpha)|\mathcal{G}| \rfloor \frac{1}{|\mathcal{G}|^{1/3}}} \leq \tau \frac{2\epsilon}{(1-\alpha)|\mathcal{G}|^{1/3}} \leq \frac{2\epsilon}{\alpha |\mathcal{G}|^{1/3}} \leq \frac{1}{2}. \tag{168}$$

Once we have shown that P_0 and P_u are probability distributions for all $u \in \mathcal{U}(Q)$, let's prove that
1) $\texttt{F\_CVaR}_\alpha(P_0, \mathbf{w}) = 0$
2) $\texttt{F\_CVaR}_\alpha(P_u, \mathbf{w}) \geq \epsilon$ for all $u \in \mathcal{U}(Q)$.

Recall that CVaR fairness is given by

$$\texttt{F\_CVaR}_\alpha(P, \mathbf{w}) = \frac{1}{1-\alpha} \max_{\sum_{g \in A \subset \mathcal{G}} w_g \leq 1-\alpha} \sum_{g \in A} w_g \Delta(P, g), \tag{169}$$

where $\Delta(g, P)$ is such that

$$\Delta(P, g) = \left| P(\hat{Y} = 1|G = g, Y = 0) - P(\hat{Y} = 1|Y = 0) \right|. \tag{170}$$

Item 1): Let's show that $\texttt{F\_CVaR}_\alpha(P_0, \mathbf{w}) = 0$. Note that for all $g \in \mathcal{G}$

$$P_0(\hat{Y} = 1|Y = 0) = \sum_{g \in \mathcal{G}} w_g P_0(\hat{Y} = 1|G = g, Y = 0) = \frac{1}{2} = P_0(\hat{Y} = 1|G = g, Y = 0), \tag{171}$$

therefore,

$$\Delta(P_0, g) = 0 \quad \forall g \in \mathcal{G}, \tag{172}$$

concluding that

$$\texttt{F\_CVaR}_\alpha(P_0, \mathbf{w}) = \frac{1}{1-\alpha} \max_{\sum_{g \in A \subset \mathcal{G}} w_g \leq 1-\alpha} \sum_{g \in A} w_g \Delta(P_0, g) = 0. \tag{173}$$

Item 2): Let's show that $\texttt{F\_CVaR}_\alpha(P_u, \mathbf{w}) \geq \epsilon$ for all $u \in \mathcal{U}(Q)$.

Note that

$$P_u(\hat{Y}=1|Y=0)=\sum_{g\in\mathcal{G}} w_g P_u(\hat{Y}=1|G=g,Y=0)=\frac{1}{2}+\sum_{g\in Q}\tau\epsilon_g u_g. \tag{174}$$

Hence, we show that $\Delta(P_u,g')$ for $g'\in Q$ is

$$\Delta(P_u,g')=\tau\left|\frac{\epsilon_{g'}u_{g'}}{w_{g'}}-\sum_{g\in Q}\epsilon_g u_g\right|. \tag{175}$$

Hence we can lower bound the CVaR fairness gap using (175).

$$\begin{aligned}
\texttt{F_CVaR}_\alpha(P_u,\mathbf{w}) &= \frac{1}{1-\alpha}\max_{\sum_{g\in A\subset\mathcal{G}} w_g\leq 1-\alpha}\sum_{g'\in A} w_{g'}\Delta(P_u,g') && (176)\\
&\geq \frac{1}{1-\alpha}\sum_{g\in Q} w_g\Delta(g) && (177)\\
&= \frac{1}{1-\alpha}\sum_{g'\in Q} w_{g'}\left|\frac{\epsilon_{g'}\tau u_{g'}}{w_{g'}}-\sum_{g\in Q}\epsilon_g\tau u_g\right| && (178)\\
&= \frac{\tau}{1-\alpha}\sum_{g'\in Q}\left|\epsilon_{g'}u_{g'}-w_{g'}\sum_{g\in Q}\epsilon_g u_g\right| && (179)\\
&\geq \frac{\tau}{1-\alpha}\sum_{g'\in Q}|\epsilon_{g'}u_{g'}|-w_{g'}\left|\sum_{g\in Q}\epsilon_g u_g\right| && (180)\\
&\geq \frac{\tau}{1-\alpha}\sum_{g'\in Q}\epsilon_{g'}-w_{g'}\sum_{g\in Q}\epsilon_g|u_g| && (181)\\
&= \frac{\tau}{1-\alpha}\sum_{g'\in Q}\epsilon_{g'}-w_{g'}\sum_{g\in Q}\epsilon_g && (182)\\
&= \frac{\tau}{1-\alpha}\sum_{g'\in Q}\epsilon_{g'}-w_{g'}\epsilon && (183)\\
&= \frac{\tau}{1-\alpha}\epsilon\left(1-\sum_{g'\in Q} w_{g'}\right) && (184)\\
&\geq \frac{\tau}{1-\alpha}\epsilon\alpha=\epsilon. && (185)
\end{aligned}$$

We then conclude that $\texttt{F\_CVaR}_\alpha(P_u,\mathbf{w})\geq\epsilon$ for all $u\in\mathcal{U}(Q)$ and from item 1 we had that $\texttt{F\_CVaR}_\alpha(P_0,\mathbf{w})=0$.

Now, let's show that, using these distributions, we can make the χ^2 distribution be small considering the mixture of distributions. Our main objective is to upper bound $\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)$ by a quantity that (i) goes to zero when the number of samples n increase, and (ii) increases when the number of groups $|\mathcal{G}|$ increases.

To bound χ^2 of a mixture, we will use the Ingster–Suslina method [43]. Hence, we assume that each u is chosen at random from a distribution π. The method ensures that

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)=\mathbb{E}_{u,u'\sim\pi}\left[\prod_{i=1}^n\sum_{z_i\in\mathcal{Z}}\frac{P_u(z_i)P_{u'}(z_i)}{P_0(z_i)}\right]-1. \tag{186}$$

Hence, we just need to compute $\sum_{z_i\in\mathcal{Z}}\frac{P_v(z_i)P_{v'}(z_i)}{P_0(z_i)}$.

$$\sum_{z_i\in\mathcal{Z}}\frac{P_u(z_i)P_{u'}(z_i)}{P_0(z_i)} \tag{187}$$

$$=\sum_{\hat{y},g}\frac{P_u(\hat{Y}=\hat{y},G=g|Y=0)P_{u'}(\hat{Y}=\hat{y},G=g|Y=0)}{P_0(\hat{Y}=\hat{y},G=g|Y=0)} \tag{188}$$

$$=\sum_{\hat{y},g\in Q}\frac{P_u(\hat{Y}=\hat{y},G=g|Y=1)P_{u'}(\hat{Y}=\hat{y},G=g|Y=0)}{P_0(\hat{Y}=\hat{y},G=g|Y=0)} \tag{189}$$

$$+\sum_{\hat{y},g\notin Q}\frac{P_u(\hat{Y}=\hat{y},G=g|Y=0)P_{u'}(\hat{Y}=\hat{y},G=g|Y=0)}{P_0(\hat{Y}=\hat{y},G=g|Y=0)} \tag{190}$$

$$=\sum_{g\in Q}\frac{\left(\frac{w_g}{2}+\tau u_g\epsilon_g\right)\left(\frac{w_g}{2}+\tau u_g^{'}\epsilon_g\right)+\left(\frac{w_g}{2}-\tau u_g\epsilon_g\right)\left(\frac{w_g}{2}-\tau u_g^{'}\epsilon_g\right)}{\frac{w_g}{2}}+\sum_{g\notin Q}\frac{2\frac{w_g}{2}\frac{w_g}{2}}{\frac{w_g}{2}} \tag{191}$$

$$=\sum_{g\in Q}\frac{\left(\frac{w_g^2}{2}+2\tau^2\epsilon_g^2u_gu_g^{'}\right)}{\frac{w_g}{2}}+\sum_{g\notin Q}\frac{2\frac{w_g}{2}\frac{w_g}{2}}{\frac{w_g}{2}} \tag{192}$$

$$=\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}+\sum_{g\in Q}w_g+\sum_{g\notin Q}w_g \tag{193}$$

$$=\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}+1. \tag{194}$$

Now, using Ingster–Suslina inequality (186) and the bound in (193), we have

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)=\mathbb{E}_{u,u'\sim\pi}\left[\prod_{i=1}^n\sum_{z_i\in\mathcal{Z}}\frac{P_u(z_i)P_{u'}(z_i)}{P_0(z_i)}\right]-1 \tag{195}$$

$$=\mathbb{E}_{u,u'\sim\pi}\left[\left(\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}+1\right)^n\right]-1 \tag{196}$$

$$\leq\mathbb{E}_{u,u'\sim\pi}\left[e^{n\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}}\right]-1 \tag{197}$$

$$\leq\mathbb{E}_{u\sim\pi}\mathbb{E}_{u'\sim\pi}\left[e^{n\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}}\right]-1. \tag{198}$$

We note that u and u' are 1-subgasian, hence, we can bound (198) by

$$\mathbb{E}_{u\sim\pi}\mathbb{E}_{u'\sim\pi}\left[e^{n\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}}\right]-1=\mathbb{E}_{u\sim\pi}\mathbb{E}_{u'\sim\pi}\left[\prod_{g\in Q}e^{n\frac{4\tau^2\epsilon_g^2u_gu_g^{'}}{w_g}}\right]-1 \tag{199}$$

$$\leq\mathbb{E}_{u\sim\pi}\left[e^{n^2\sum_{g\in Q}\frac{16\tau^4\epsilon_g^4u_g^2}{w_g^2}}\right]-1 \tag{200}$$

$$\leq e^{n^2\sum_{g\in Q}\frac{16\tau^4\epsilon_g^4}{w_g^2}}-1, \tag{201}$$

where (199) comes from (198), (200) comes from the fact that u' is 1-subgausian, and (200) comes from the fact that $|u_g|=1$.

Now, by plugging in ϵ_g (166) in (201) we conclude that

$$\mathbb{E}_{u\sim\pi}\mathbb{E}_{u'\sim\pi}\left[e^{n\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2 u_g u_g'}{w_g}}\right]-1\leq e^{n^2\sum_{g\in Q}\frac{16\tau^4\epsilon_g^4}{w_g^2}}-1\leq e^{\frac{16\tau^4 n^2\epsilon^4}{\left(\sum_{g\in Q}w_g^{2/3}\right)^3}}-1 \tag{202}$$

But, from the definition of Q, i.e., $\sum_{g\in Q}w_g^{2/3}=\lfloor(1-\alpha)|\mathcal{G}|\rfloor\frac{1}{|\mathcal{G}|^{1/3}}\geq\frac{3(1-\alpha)}{8}\sum_{g\in\mathcal{G}}w_g^{2/3}$. Hence, plugging (202) in (202) we have that

$$\mathbb{E}_{u\sim\pi}\mathbb{E}_{u'\sim\pi}\left[e^{n\sum_{g\in Q}\frac{4\tau^2\epsilon_g^2 u_g u_g'}{w_g}}\right]-1\leq e^{n^2\sum_{g\in Q}\frac{16\tau^4\epsilon_g^4}{w_g^2}}-1\leq e^{\frac{1024\tau^4 n^2\epsilon^4}{\left((1-\alpha)\sum_{g\in\mathcal{G}}w_g^{2/3}\right)^3}}-1. \tag{203}$$

Concluding the proof of the lemma by combining the inequality in (198) with (203) and getting

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)\leq e^{\frac{1024\tau^4 n^2\epsilon^4}{\left((1-\alpha)\sum_{g\in\mathcal{G}}w_g^{2/3}\right)^3}}-1 \tag{204}$$

$$\leq e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4\left(\sum_{g\in\mathcal{G}}w_g^{2/3}\right)^3}}-1. \tag{205}$$

□

Theorem 3: (Minimum Sample Complexity). For CVaR fairness metrics $\texttt{F\_CVaR}_\alpha(\mathbf{w})$ with quality of service function L and measuring the average quality of service using the distribution P such that $(L,P)\in\{(L_{EO},P_{EO}),(L_{SP},P_{SP})\}$. Consider the hypothesis regions $\mathcal{P}_0$ (3) and $\mathcal{P}_1(\epsilon)$ (4). Assume that $\mathbf{w}$ is a uniform distribution over all groups, i.e., $\mathbf{w}=\left(\frac{1}{|\mathcal{G}|},...,\frac{1}{|\mathcal{G}|}\right)$. For all $\alpha\in[0,1)$, $\frac{\alpha|\mathcal{G}|^{1/3}}{4}\geq\epsilon>0$, the minimum probability of error in the ϵ-test using n i.i.d. samples from P to distinguish between $\texttt{F\_CVaR}_\alpha(\mathbf{w})=0$ and $\texttt{F\_CVaR}_\alpha(\mathbf{w})\geq\epsilon$ is lower bounded by

$$2\inf_\psi P_{\text{error}}\geq 1-\frac{1}{2}\sqrt{e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}}-1}, \tag{45}$$

then n samples are needed to perform an ϵ-test with with probability of error smaller than 0.01 where n is such that

$$n=\Omega\left(\frac{\alpha^2\sqrt{|\mathcal{G}|}}{(1-\alpha)^{1/2}\epsilon^2}\right). \tag{46}$$

Proof: From Lemma 1 we know that there exist a sequence of distributions $P_0\in\mathcal{P}_0(\epsilon)$ and $\{P_u\}_{u\in\mathcal{U}}\in\mathcal{P}_1(\epsilon)$ indexed by elements of a set $\mathcal{U}$ such that if $u\in\mathcal{U}$ we have that

$$\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)\leq e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}}-1. \tag{206}$$

On the other hand, by the generalized Le Cam lemma [41], we have that

$$2\inf_\psi P_{\text{error}}\geq 1-\text{TV}\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right) \tag{207}$$

$$\geq 1-\frac{1}{2}\sqrt{\chi^2\left(\mathbb{E}_u\left[P_u^n\right]\middle\|P_0^n\right)} \tag{208}$$

Plugging (206) in (208), we have that:

$$2\inf_\psi P_{\text{error}}\geq 1-\frac{1}{2}\sqrt{e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}}-1}. \tag{209}$$

Therefore, if the probability of error is smaller than 0.01 we need to ensure that:

$$0.02 \geq 1 - \frac{1}{2}\sqrt{e^{\frac{1024(1-\alpha)n^2\epsilon^4}{\alpha^4|\mathcal{G}|}} - 1}. \tag{210}$$

From where we conclude that

$$n = \Omega\left(\frac{\alpha^2\sqrt{|\mathcal{G}|}}{(1-\alpha)^{1/2}\epsilon^2}\right), \tag{211}$$

which is the desired result. □

Lemma 11 ($\texttt{F\_CVaR}_\alpha$ *vs. CVaR*$_\beta$*):* Let $\sum_{g\in Q_\alpha} w_g = 1-\alpha$, and any two groups $g \neq g'$ are such that $\Delta(g) \neq \Delta(g')$. Then, we can define $\beta = 1-\alpha$ and conclude that

$$\text{CVaR}_\beta(\Delta(g)) = \texttt{F\_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})) \tag{212}$$

Proof: First, notice that $g \in Q_\alpha$ if and only if $\Delta(g) \geq q$ Hence, taking $\beta = 1-\alpha$ we conclude that $q_\beta = \min_{g\in Q_\alpha} \Delta(g)$, then

$$\begin{aligned}
\text{CVaR}_\beta(\Delta(g)) &= \mathbb{E}[\Delta(g)|\Delta(g) \geq q_\beta] && (213)\\
&= \frac{\sum_{g;\Delta(g)\geq\beta} w_g\Delta(g)}{\sum_{g;\Delta(g)\geq\beta} w_g} && (214)\\
&= \frac{\sum_{g\in Q_\alpha} w_g\Delta(g)}{\sum_{g\in Q_\alpha} w_g} && (215)\\
&= \frac{\sum_{g\in Q_\alpha} w_g\Delta(g)}{1-\alpha} && (216)\\
&= \texttt{F_CVaR}_\alpha(\mathbf{w}; P, L, \overline{L}(\mathbf{w})). && (217)
\end{aligned}$$

□