Certifying Some Distributional Fairness with Subpopulation Decomposition

Mintong Kang *
UIUC
mintong2@illinois.edu

Linyi Li *
UIUC
linyi2@illinois.edu

Maurice Weber
ETH Zurich
webermau@inf.ethz.ch

Yang Liu
UC Santa Cruz
yangliu@ucsc.edu

Ce Zhang
ETH Zurich
ce.zhang@inf.ethz.ch

Bo Li
UIUC
lbo@illinois.edu

Abstract

Extensive efforts have been made to understand and improve the fairness of machine learning models based on different fairness measurement metrics, especially in high-stakes domains such as medical insurance, education, and hiring decisions. However, there is a lack of *certified fairness* on the end-to-end performance of an ML model. In this paper, we first formulate the certified fairness of an ML model trained on a given data distribution as an optimization problem based on the model performance loss bound on a fairness constrained distribution, which is within bounded distributional distance with the training distribution. We then propose a general fairness certification framework and instantiate it for both sensitive shifting and general shifting scenarios. In particular, we propose to solve the optimization problem by decomposing the original data distribution into analytical subpopulations and proving the convexity of the sub-problems to solve them. We evaluate our certified fairness on six real-world datasets and show that our certification is tight in the sensitive shifting scenario and provides non-trivial certification under general shifting. Our framework is flexible to integrate additional non-skewness constraints and we show that it provides even tighter certification under different real-world scenarios. We also compare our certified fairness bound with adapted existing distributional robustness bounds on Gaussian data and demonstrate that our method is significantly tighter.

1 Introduction

As machine learning (ML) has become ubiquitous [24, 18, 5, 11, 8, 13], fairness of ML have attracted a lot of attention from different perspectives. For instance, some automated hiring systems are biased towards males due to gender imbalanced training data [3]. Different approaches have been proposed to improve ML fairness, such as regularized training [16, 22, 26, 30], disentanglement [12, 28, 40], duality [44], low-rank matrix factorization [34], and distribution alignment [4, 29, 53].

In addition to existing approaches that *evaluate* fairness, it is important and challenging to provide *certification* for ML fairness. Recent studies have explored the certified fair *representation* of ML [39, 4, 36]. However, there lacks certified fairness on the *predictions* of an end-to-end ML model trained on an arbitrary data distribution. In addition, current fairness literature mainly focuses on training an ML model on a potentially (im)balanced distribution and evaluate its performance in a target domain measured by existing statistical fairness definitions [17, 20]. Since in practice these selected target domains can encode certain forms of unfairness of their own (e.g., sampling bias), the evaluation would be more informative if we can evaluate and certify fairness of an ML model

*The first two authors contributed equally.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

on an *objective* distribution. Taking these factors into account, in this work, we aim to provide the first definition of *certified fairness* given an ML model and a training distribution by bounding its end-to-end performance on an objective, *fairness constrained distribution*. In particular, we define *certified fairness* as the worst-case upper bound of the ML prediction loss on a fairness constrained test distribution $\mathcal{Q}$, which is within a bounded distance to the training distribution $\mathcal{P}$. For example, for an ML model of crime rate prediction, we can define the model performance as the expected loss within a specific age group. Suppose the model is deployed in a fair environment that does not deviate too much from the training, our fairness certificate can guarantee that the loss of crime rate prediction for a particular age group is upper bounded, which is an indicator of model's fairness.

We mainly focus on the base rate condition as the fairness constraint for $\mathcal{Q}$. We prove that our certified fairness based on a base rate constrained distribution will imply other fairness metrics, such as demographic parity (DP) and equalized odds (EO). Moreover, our framework is flexible to integrate other fairness constraints into $\mathcal{Q}$. We consider two scenarios: (1) *sensitive shifting* where only the joint distribution of sensitive attribute and label can be changed when optimizing $\mathcal{Q}$; and (2) *general shifting* where everything including the conditioned distribution of non-sensitive attributes can be changed. We then propose an effective *fairness certification framework* to compute the certificate.

In our fairness certification framework, we first formulate the problem as constrained optimization, where the fairness constrained distribution is encoded by base rate constraints. Our key technique is to decompose both training and the fairness constrained test distributions to several subpopulations based on sensitive attributes and target labels, which can be used to encode the base rate constraints. With such a decomposition, in sensitive shifting, we can decompose the distance constraint to subpopulation ratio constraints and prove the transformed low-dimensional optimization problem is convex and thus efficiently solvable. In general shifting case, we propose to solve it based on divide and conquer: we first partition the feasible space into different subpopulations, then optimize the density (ratio) of each subpopulation, apply relaxation on each subpopulation as a sub-problem, and finally prove the convexity of the sub-problems with respect to other low-dimensional variables. Our framework is applicable for any black-box ML models and any distributional shifts bounded by the Hellinger distance, which is a type of f-divergence studied in the literature [47, 14, 7, 25, 15].

To demonstrate the effectiveness and tightness of our framework, we evaluate our fairness bounds on *six* real-world fairness related datasets [3, 2, 19, 48]. We show that our certificate is tight under different scenarios. In addition, we verify that our framework is flexible to integrate additional constraints on $\mathcal{Q}$ and evaluate the certified fairness with additional non-skewness constraints, with which our fairness certificate is tighter. Finally, as the first work on certifying fairness of an end-to-end ML model, we adapt existing distributional robustness bound [43] for comparison to provide more intuition. Note that directly integrating the fairness constraint to the existing distributional robustness bound is challenging, which is one of the main contributions for our framework. We show that with the fairness constraints and our effective solution, our bound is strictly tighter.

Technical Contributions. In this work, we take the first attempt towards formulating and computing the *certified fairness* on an end-to-end ML model, which is trained on a given distribution. We make contributions on both theoretical and empirical fronts.

1. We formulate the *certified fairness* of an end-to-end ML model trained on a given distribution $\mathcal{P}$ as the worst-case upper bound of its prediction loss on a fairness constrained distribution $\mathcal{Q}$, which is within bounded distributional distance with $\mathcal{P}$.
2. We propose an effective fairness certification framework that simulates the problem as constrained optimization and solve it by decomposing the training and fairness constrained test distributions into subpopulations and proving the convexity of each sub-problem to solve it.
3. We evaluate our certified fairness on six real-world datasets to show its tightness and scalability. We also show that with additional distribution constraints on $\mathcal{Q}$, our certification would be tighter.
4. We show that our bound is strictly tighter than adapted distributional robustness bound on Gaussian dataset due to the added fairness constraints and our effective optimization approach.

Related Work Fairness in ML can be generally categorized into individual fairness and group fairness. Individual fairness guarantees that similar inputs should lead to similar outputs for a model and it is analyzed with optimization approaches [49, 33] and different types of relaxations [21]. Group fairness indicates to measure the *independence* between the sensitive features and model prediction, the *separation* which means that the sensitive features are statistically independent of model prediction given the target label, and the *sufficiency* which means that the sensitive features are

statistically independent of the target label given the model prediction [27]. Different approaches are proposed to analyze group fairness via static analysis [46], interactive computation [41], and probabilistic approaches [1, 10, 6]. In addition, there is a line of work trying to certify the *fair representation* [39, 4, 36]. In [9], the authors have provided bounds for how group fairness transfers subject to bounded distribution shift. Our certified fairness differs from existing work from three perspectives: 1) we provide fairness certification considering the end-to-end model performance instead of the representation level, 2) we define and certify fairness based on a fairness constrained distribution which implies other fairness notions, and 3) our certified fairness can be computed for *any* black-box models trained on an arbitrary given data distribution.

2 Certified Fairness Based on Fairness Constrained Distribution

In this section, we first introduce preliminaries, and then propose the definition of *certified fairness* based on a bounded fairness constrained distribution, which to the best of our knowledge is the first formal fairness certification on end-to-end model prediction. We also show that our proposed certified fairness relates to established fairness definitions in the literature.

Notations. We consider the general classification setting: we denote by $\mathcal{X}$ and $\mathcal{Y} = [C]$ the feature space and labels, $[C] := \{1, 2, \cdots, C\}$. $h_\theta \colon \mathcal{X} \to \Delta^{|\mathcal{Y}|}$ represents a mapping function parameterized with $\theta \in \Theta$, and $\ell \colon \Delta^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}_+$ is a non-negative loss function such as cross-entropy loss. Within feature space $\mathcal{X}$, we identify a *sensitive* or *protected attribute* $\mathcal{X}_s$ that takes a finite number of values: $\mathcal{X}_s := [S]$, i.e., for any $X \in \mathcal{X}$, $X_s \in [S]$.

Definition 1 (Base Rate)**.** Given a distribution $\mathcal{P}$ supported over $\mathcal{X} \times \mathcal{Y}$, the base rate for sensitive attribute value $s \in [S]$ with respect to label $y \in [C]$ is $b^{\mathcal{P}}_{s,y} = \Pr_{(X,Y)\sim\mathcal{P}}[Y = y \,|\, X_s = s]$.

Given the definition of base rate, we define a *fair base rate distribution* (in short as *fair distribution*).

Definition 2 (Fair Base Rate Distribution)**.** A distribution $\mathcal{P}$ supported over $\mathcal{X} \times \mathcal{Y}$ is a fair base rate distribution if and only if for any label $y \in [C]$, the base rate $b^{\mathcal{P}}_{s,y}$ is equal across all $s \in [S]$, i.e., $\forall i \in [S], \forall j \in [S], b^{\mathcal{P}}_{i,y} = b^{\mathcal{P}}_{j,y}$.

Remark. In the literature, the concepts of fairness are usually directly defined at the model prediction level, where the criterion is whether the model prediction is fair against individual attribute changes [39, 36, 50] or fair at population level [54]. In this work, to certify the fairness of model prediction, we define a fairness constrained distribution on which we will certify the model prediction (e.g., bound the prediction error), rather than relying on the empirical fairness evaluation. In particular, we first define the fairness constrained distribution through the lens of base rate parity, i.e., the probability of being any class should be independent of sensitive attribute values, and then define the certified fairness of a given model based on its performance on the fairness constrained distribution as we will show next.

The choice of focusing on fair base rate may look restrictive but its definition aligns very well with the celebrated fairness definition Demographic Parity [51], which promotes that $\Pr[h_\theta(X) = 1|X_s = i] = \Pr[h_\theta(X) = 1|X_s = j]$. In this case, the prediction performance of h_θ on $\mathcal{Q}$ with fair base rate will relate directly to $\Pr[h_\theta(X) = 1|X_s = i]$. Secondly, under certain popular data generation process, the base rate sufficiently encodes the differences in distributions and a fair base rate will imply a homogeneous (therefore equal or "fair") distribution over X, Y: consider when $\Pr(X|Y = y, X_s = i)$ is the same across different group X_s. Then $\Pr(X, Y|X_s = i)$ is simply a linear combination of basis distributions $\Pr(X|Y = y, X_s = i)$, and the difference between different groups' joint distribution of X, Y is fully characterized by the difference in base rate $\Pr(Y = y|X_s)$. This assumption will greatly enable trackable analysis and is not an uncommon modeling choice in the recent discussion of fairness when distribution shifts [52, 37].

2.1 Certified Fairness

Now we are ready to define the fairness certification based on the optimized fairness constrained distribution. We define the certification under two data generation scenarios: *general shifting* and *sensitive shifting*. In particular, consider the data generative model $\Pr(X_o, X_s, Y) = \Pr(Y)\Pr(X_s|Y)\Pr(X_o|Y, X_s)$, where X_o and X_s represent the non-sensitive and sensitive features, respectively. If all three random variables on the RHS are allowed to change, we call it *general shifting*; if both $\Pr(Y)$ and $\Pr(X_s|Y)$ are allowed to change to ensure the fair base rate (Def. 2)

while $\Pr(X_o|Y, X_s)$ is the same across different groups, we call it *sensitive shifting*. In Section 3 we will introduce our certification framework for both scenarios.

Problem 1 (Certified Fairness with General Shifting)**.** *Given a training distribution $\mathcal{P}$ supported on $\mathcal{X} \times \mathcal{Y}$, a model $h_\theta(\cdot)$ trained on $\mathcal{P}$, and distribution distance bound $\rho > 0$, we call $\bar{\ell} \in \mathbb{R}$ a* fairness certificate *with general shifting, if $\bar{\ell}$ upper bounds*

$$\max_{\mathcal{Q}} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)] \quad \text{s.t.} \quad \text{dist}(\mathcal{P}, \mathcal{Q}) \leq \rho, \quad \mathcal{Q} \textit{ is a fair distribution},$$

where $\text{dist}(\cdot, \cdot)$ *is a predetermined distribution distance metric.*

In the above definition, we define the fairness certificate as the upper bound of the model's loss among all fair base rate distributions $\mathcal{Q}$ within a bounded distance from $\mathcal{P}$. Besides the bounded distance constraint $\text{dist}(\mathcal{P}, \mathcal{Q}) \leq \rho$, there is no other constraint between $\mathcal{P}$ and $\mathcal{Q}$ so this satisfies "*general shifting*". This bounded distance constraint, parameterized by a tunable parameter ρ, ensures that the test distribution should not be too far away from the training. In practice, the model h_θ may represent a DNN whose complex analytical forms would pose challenges for solving Problem 1. As a result, as we will show in Equation (2) we can query some statistics of h_θ trained on $\mathcal{P}$ as constraints to characterize h_θ, and thus compute the upper bound certificate.

The feasible region of optimization problem 1 might be empty if the distance bound ρ is too small, and thus we cannot provide fairness certification in this scenario, indicating that there is no nearby fair distribution and thus the fairness of the model trained on the highly "unfaired" distribution is generally low. In other words, if the training distribution $\mathcal{P}$ is unfair (typical case) and there is no feasible fairness constrained distribution $\mathcal{Q}$ within a small distance to $\mathcal{P}$, fairness cannot be certified.

This definition follows the intuition of typical real-world scenarios: The real-world training dataset is usually biased due to the limitation in data curation and collection processes, which causes the model to be unfair. Thus, when the trained models are evaluated on the real-world fairness constrained test distribution or ideal fair distribution, we hope that the model does not encode the training bias which would lead to low test performance. That is to say, the model performance on fairness constrained distribution is indeed a witness of the model's intrinsic fairness.

We can further constrain that the subpopulation of $\mathcal{P}$ and $\mathcal{Q}$ parameterized by X_s and Y does not change, which results in the following "*sensitive shifting*" fairness certification.

Problem 2 (Certified Fairness with Sensitive Shifting)**.** *Under the same setting as Problem 1, we call $\bar{\ell}$ a fairness certificate against sensitive shifting, if $\bar{\ell}$ upper bounds*

$$\begin{aligned} &\max_{\mathcal{Q}} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)] \\ \text{s.t.} \quad &\text{dist}(\mathcal{P}, \mathcal{Q}) \leq \rho, \quad \mathcal{P}_{s,y} = \mathcal{Q}_{s,y}\, \forall s \in [S], y \in [C], \quad \mathcal{Q} \textit{ is a fair distribution}, \end{aligned}$$

where $\mathcal{P}_{s,y}$ and $\mathcal{Q}_{s,y}$ are the subpopulations of $\mathcal{P}$ and $\mathcal{Q}$ on the support $\{(X, Y) : X \in \mathcal{X}, X_s = s, Y = y\}$ respectively, and $\text{dist}(\cdot, \cdot)$ *is a predetermined distribution distance metric.*

The definition adds an additional constraint between $\mathcal{P}$ and $\mathcal{Q}$ that each subpopulation, partitioned by the sensitive attribute X_s and label Y, does not change. This constraint corresponds to the scenario where the distribution shifting between training and test distributions only happens on the proportions of different sensitive attributes and labels, and within each subpopulation the shifting is negligible.

In addition, to model the real-world test distribution, we may further request that the test distribution $\mathcal{Q}$ is not too skewed regarding the sensitive attribute X_s by adding constraint (1). We will show that this constraint can also be integrated into our fairness certification framework flexibly in Section 4.3.

$$\forall i \in [S], \forall j \in [S], \left| \Pr_{(X,Y)\sim\mathcal{Q}}[X_s = i] - \Pr_{(X,Y)\sim\mathcal{Q}}[X_s = j] \right| \leq \Delta_S. \tag{1}$$

Connections to Other Fairness Measurements. Though not explicitly stated, our goal of certifying the performance on a fair distribution $\mathcal{Q}$ relates to certifying established fairness definitions in the literature. Consider the following example: Suppose Problem 2 is feasible and returns a classifier h_θ that achieves certified fairness per group and per label class $\bar{l} := \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) \neq Y | Y = y, X_s = i] \leq \epsilon$ on $\mathcal{Q}$. We will then have the following proposition:

Proposition 1. *h_θ achieves ϵ-Demographic Parity (DP) [51] and ϵ-Equalized Odds (EO) [18]:*

- *ϵ-DP:* $|\Pr_{\mathcal{Q}}[h_\theta(X) = 1 | X_s = i] - \Pr_{\mathcal{Q}}[h_\theta(X) = 1 | X_s = j]| \leq \epsilon,\ \forall i, j.$

- *ϵ-EO:* $|\Pr_{\mathcal{Q}}[h_\theta(X)=1|Y=y, X_s=i] - \Pr_{\mathcal{Q}}[h_\theta(X)=1|Y=y, X_s=j]| \leq \epsilon, \forall y, i, j$.

Remark. The detailed proof is omitted to appendix C.1. (1) When $\epsilon = 0$, Proposition 1 can guarantee perfect DP and EO simultaneously. We achieve so because we evaluate with a fair distribution $\mathcal{Q}$, where "fair distribution" stands for "equalized base rate" and according to [23, Theorem 1.1, page 5] both DP and EO are achievable for this fair distribution. This observation in fact motivated us to identify the fair distribution $\mathcal{Q}$ for the evaluation since it is this fair distribution that allows the fairness measures to hold at the same time. Therefore, another way to interpret our framework is: given a model, we provide a framework that certifies worst-case "unfairness" bound in the context where perfect fairness is achievable. Such a worse-case bound serves as the gap to a perfectly fair model and could be a good indicator of the model's fairness level. (2) In practice, ϵ is not necessarily zero. Therefore, Proposition 1 only provides an upper lower bound of DP and EO, namely ϵ-DP and ϵ-EO, instead of absolute DP and EO. The approximate fairness guarantee renders our results more general. Meanwhile, there is a higher flexiblity in simultaneously satisfying approximate fairness metrics (for example when DP $= 0$, but EO = ϵ, which is plausible for a proper range of epsilon, regardless of the distribution $\mathcal{Q}$ being fair or not). But again, similar to (1), ϵ-DP and ϵ-EO can be achieved at the same time easily since the test distribution satisfies base rate parity.

The bounds in Proposition 1 are tight. Consider the distribution $\mathcal{Q}$ with binary classes and binary sensitive attributes (i.e., $Y, X_s \in \{0,1\}$). When the distribution $\mathcal{Q}$ and classifier h_θ satisfy the conditions that $\Pr_{\mathcal{Q}}[h_\theta(X) \neq Y|Y=0, X_s=0] = \epsilon, \Pr_{\mathcal{Q}}[h_\theta(X) \neq Y|Y=0, X_s=1] = 0$ and $\Pr_{\mathcal{Q}}[Y=0] = 1, \Pr_{\mathcal{Q}}[Y=1] = 0$, the bounds in Proposition 1 are tight. From $\Pr_{\mathcal{Q}}[Y=0] = 1, \Pr_{\mathcal{Q}}[Y=1] = 0$, we can observe that ϵ-DP is equivalent to ϵ-EO. From $\Pr_{\mathcal{Q}}[h_\theta(X) \neq Y|Y = 0, X_s = 0] = \epsilon, \Pr_{\mathcal{Q}}[h_\theta(X) \neq Y|Y=0, X_s=1] = 0$ and $\Pr_{\mathcal{Q}}[h_\theta(X) \neq Y|Y=0, X_s=i] = \Pr_{\mathcal{Q}}[h_\theta(X)=1|Y=0, X_s=i]$ for $i \in \{0,1\}$, we know that ϵ-EO holds with tightness since $|\Pr_{\mathcal{Q}}[h_\theta(X)=1|Y=0, X_s=0] - \Pr_{\mathcal{Q}}[h_\theta(X)=1|Y=0, X_s=1]| = \epsilon$. To this point, we show that both bounds in Proposition 1 are tight.

3 Fairness Certification Framework

We will introduce our fairness certification framework which efficiently computes the fairness certificate defined in Section 2.1. We first introduce our framework for *sensitive shifting* (Problem 2) which is less complex and shows our core methodology, then *general shifting* case (Problem 1).

Our framework focuses on using the Hellinger distance to bound the distributional distance in Problems 1 and 2. The Hellinger distance $H(\mathcal{P}, \mathcal{Q})$ is defined in Def. 3 (in Appendix B.1). The Hellinger distance has some nice properties, e.g., $H(\mathcal{P}, \mathcal{Q}) \in [0,1]$, and $H(\mathcal{P}, \mathcal{Q}) = 0$ if and only if $\mathcal{P} = \mathcal{Q}$ and the maximum value of 1 is attained when $\mathcal{P}$ and $\mathcal{Q}$ have disjoint support. The Hellinger distance is a type of f-divergences which are widely studied in ML distributional robustness literature [47, 14] and in the context of distributionally robust optimization [7, 25, 15]. Also, using Hellinger distance enables our certification framework to generalize to *total variation distance (or statistic distance)* $\delta(\mathcal{P}, \mathcal{Q})$[2] directly with the connection, $H^2(\mathcal{P}, \mathcal{Q}) \leq \delta(\mathcal{P}, \mathcal{Q}) \leq \sqrt{2}H(\mathcal{P}, \mathcal{Q})$ ([45], Equation 1). We leave the extension of our framework to other distance metrics as future work.

3.1 Core Idea: Subpopulation Decomposition

The core idea in our framework is (finite) subpopulation decomposition. Consider a generic optimization problem for computing the loss upper bound on a constrained test distribution $\mathcal{Q}$, given training distribution $\mathcal{P}$ and trained model $h_\theta(\cdot)$, we first characterize model $h_\theta(\cdot)$ based on some statistics, e.g., mean and variance for loss of the model: $h_\theta(\cdot)$ satisfies $e_j(\mathcal{P}, h_\theta) \leq v_j$, $1 \leq j \leq L$. Then we characterize the properties (e.g., fair base rate) of the test distribution $\mathcal{Q}$: $g_j(\mathcal{Q}) \leq u_j$, $1 \leq j \leq M$. As a result, we can upper bound the loss of $h_\theta(\cdot)$ on $\mathcal{Q}$ as the following optimization:

$$\max_{\mathcal{Q},\theta} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)] \quad \text{s.t.} \quad H(\mathcal{P}, \mathcal{Q}) \leq \rho, \quad e_j(\mathcal{P}, h_\theta) \leq v_j \, \forall j \in [L], \quad g_j(\mathcal{Q}) \leq u_j \, \forall j \in [M]. \tag{2}$$

Now we decompose the space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ to N partitions: $\mathcal{Z} := \biguplus \mathcal{Z}_i$, where $\mathcal{Z}$ is the support of both $\mathcal{P}$ and $\mathcal{Q}$. Then, we denote $\mathcal{P}$ conditioned on $\mathcal{Z}_i$ by $\mathcal{P}_i$ and similarly $\mathcal{Q}$ conditioned on $\mathcal{Z}_i$ by $\mathcal{Q}_i$. As a result, we can write $\mathcal{P} = \sum_{i\in[N]} p_i \mathcal{P}_i$ and $\mathcal{Q} = \sum_{i\in[N]} q_i \mathcal{Q}_i$. Since $\mathcal{P}$ is known, p_i's are known. In contrast, both $\mathcal{Q}_i$ and q_i's are optimizable. Our key observation is that

$$H(\mathcal{P}, \mathcal{Q}) \leq \rho \iff 1 - \rho^2 - \sum_{i=1}^{N} \sqrt{p_i q_i}(1 - H(\mathcal{P}_i, \mathcal{Q}_i)^2) \leq 0 \tag{3}$$

[2] $\delta(\mathcal{P}, \mathcal{Q}) = \sup_{A\in\mathcal{F}} |\mathcal{P}(A) - \mathcal{Q}(A)|$ where $\mathcal{F}$ is a σ-algebra of subsets of the sample space Ω.

which leads to the following theorem.

Theorem 1. *The following constrained optimization upper bounds Equation* (2)*:*

$$\max_{\mathcal{Q}_i, q_i, \rho_i, \theta} \quad \sum_{i=1}^{N} q_i \mathbb{E}_{(X,Y)\sim\mathcal{Q}_i}[\ell(h_\theta(X), Y)] \tag{4a}$$

$$\text{s.t.} \quad 1 - \rho^2 - \sum_{i=1}^{N} \sqrt{p_i q_i}(1 - \rho_i^2) \leq 0, \tag{4b}$$

$$H(\mathcal{P}_i, \mathcal{Q}_i) \leq \rho_i \quad \forall i \in [N], \quad \sum_{i=1}^{N} q_i = 1, \quad q_i \geq 0 \quad \forall i \in [N], \quad \rho_i \geq 0 \quad \forall i \in [N], \tag{4c}$$

$$e'_j(\{\mathcal{P}_i\}_{i\in[N]}, \{p_i\}_{i\in[N]}, h_\theta) \leq v'_j \, \forall j \in [L], \quad g'_j(\{\mathcal{Q}_i\}_{i\in[N]}, \{q_i\}_{i\in[N]}) \leq u'_j \, \forall j \in [M], \tag{4d}$$

if $e_j(\mathcal{P}, h_\theta) \leq v_j$ *implies* $e'_j(\{\mathcal{P}_i\}_{i\in[N]}, \{p_i\}_{i\in[N]}, h_\theta) \leq v'_j$ *for any* $j \in [L]$*, and* $g_j(\mathcal{Q}) \leq u_j$ *implies* $g'_j(\{\mathcal{Q}_i\}_{i\in[N]}, \{q_i\}_{i\in[N]}) \leq u'_j$ *for any* $j \in [M]$*.*

In Problem 2, the challenge is to deal with the fair base rate constraint. Our core technique in Thm. 1 is subpopulation decomposition. At a high level, thanks to the disjoint support among different subpopulations, we get Equation (3). This equation gives us an equivalence relationship between distribution-level (namely, $\mathcal{P}$ and $\mathcal{Q}$) distance constraint and subpopulation-level (namely, $\mathcal{P}_i$'s and $\mathcal{Q}_i$'s) distance constraint. As a result, we can rewrite the original problem (2) using sub-population as decision variables as in Equation (4b) and then imposing the unity constraint (Equation (4c)) to get Thm. 1. We provide a detailed proof in Appendix C.2. Although the optimization problem (Equation (4)) may look more complicated then the original Equation (2), this optimization simplifies the challenging fair base rate constraint, allows us to upper bound each subpopulation loss $\mathbb{E}_{(X,Y)\sim\mathcal{Q}_i}[\ell(h_\theta(X), Y)]$ individually, and hence makes the whole optimization tractable.

3.2 Certified Fairness with Sensitive Shifting

For the sensitive shifting case, we instantiate Thm. 1 and obtain the following fairness certificate.

Theorem 2. *Given a distance bound* $\rho > 0$*, the following constrained optimization, which is* ***convex****, when feasible, provides a* ***tight*** *fairness certificate for Problem 2:*

$$\max_{k_s, r_y} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y E_{s,y}, \quad \text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C],$$
$$1 - \rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y} \leq 0,$$

where $E_{s,y} := \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$ *and* $p_{s,y} := \Pr_{(X,Y)\in\mathcal{P}}[X_s = s, Y = y]$ *are constants.*

Proof sketch. We decompose distribution $\mathcal{P}$ and $\mathcal{Q}$ to $\mathcal{P}_{s,y}$'s and $\mathcal{Q}_{s,y}$'s according to their sensitive attribute and label values. In sensitive shifting, $\Pr(X_o|Y, X_s)$ is fixed, i.e., $\mathcal{P}_{s,y} = \mathcal{Q}_{s,y}$, which means $\mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X), Y)] = E_{s,y}$ and $\rho_{s,y} = H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) = 0$. We plug these properties into Thm. 1. Then, denoting $q_{s,y}$ to $\Pr_{(X,Y)\sim\mathcal{Q}}[X_s = s, Y = y]$, we can represent the fairness constraint in Def. 2 as $q_{s_0,y_0} = \left(\sum_{s=1}^{S} q_{s,y_0}\right)\left(\sum_{y=1}^{C} q_{s_0,y}\right)$ for any $s_0 \in [S]$ and $y_0 \in [C]$. Next, we parameterize $q_{s,y}$ with $k_s r_y$. Such parameterization simplifies the fairness constraint and allow us to prove the convexity of the resulting optimization. Since all the constraints are encoded equivalently, the problem formulation provides a tight certification. Detailed proof in Appendix C.3. □

As Thm. 2 suggests, we can exploit the expectation information $E_{s,y} = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$ and density information $p_{s,y} = \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y]$ of each $\mathcal{P}$'s subpopulation to provide a tight fairness certificate in sensitive shifting. The convex optimization problem with $(S+C)$ variables can be efficiently solved by off-the-shelf packages.

3.3 Certified Fairness with General Shifting

For the general shifting case, we leverage Thm. 1 and the parameterization trick $q_{s,y} := k_s r_y$ used in Thm. 2 to reduce Problem 1 to the following constrained optimization.

Lemma 3.1. *Given a distance bound $\rho > 0$, the following constrained optimization, when feasible, provides a **tight** fairness certificate for Problem 1:*

$$\max_{k_s, r_y, \mathcal{Q}, \rho_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X), Y)] \tag{6a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C], \tag{6b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y}(1 - \rho_{s,y}^2) \geq 1 - \rho^2 \tag{6c}$$

$$H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \rho_{s,y} \quad \forall s \in [S], y \in [C], \tag{6d}$$

where $p_{s,y} := \Pr_{(X,Y)\in\mathcal{P}}[X_s = s, Y = y]$ is a fixed constant. The $\mathcal{P}_{s,y}$ and $\mathcal{Q}_{s,y}$ are the subpopulations of $\mathcal{P}$ and $\mathcal{Q}$ on the support $\{(X,Y) : X \in \mathcal{X}, X_s = s, Y = y\}$ respectively.

Proof sketch. We show that Equation (6b) ensures a parameterization of $q_{s,y} = \Pr_{(X,Y)\in\mathcal{Q}}[X_s = s, Y = y]$ that satisfies fairness constraints on $\mathcal{Q}$. Then, leveraging Thm. 1 we prove that the constrained optimization provides a fairness certificate. Since all the constraints are either kept or equivalently encoded, this resulting certification is *tight*. Detailed proof in Appendix C.4. □

Now the main obstacle is to solve the non-convex optimization in Problem 6. Here, as the first step, we upper bound the loss of $h_\theta(\cdot)$ within each shifted subpopulation $\mathcal{Q}_{s,y}$, i.e., upper bound $\mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X), Y)]$ in Equation (6a), by Thm. 4 in Appendix B.2 [47]. Then, we apply variable transformations to make some decision variables convex. For the remaining decision variables, we observe that they are non-convex but bounded. Hence, we propose the technique of grid-based sub-problem construction. Concretely, we divide the feasible region regarding non-convex variables into small grids and consider the optimization problem in each region individually. For each sub-problem, we relax the objective by pushing the values of non-convex variables to the boundary of the current grid and then solve the convex optimization sub-problems. Concretely, the following theorem states our computable certificate for Problem 1, with detailed proof in Appendix C.5.

Theorem 3. *If for any $s \in [S]$ and $y \in [Y]$, $H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \bar{\gamma}_{s,y}$ and $0 \leq \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \ell(h_\theta(X), Y) \leq M$, given a distance bound $\rho > 0$, for any region granularity $T \in \mathbb{N}_+$, the following expression provides a fairness certificate for Problem 1:*

$$\bar{\ell} = \max_{\{i_s\in[T]:s\in[S]\},\{j_y\in[T]:y\in[C]\}} \mathbf{C}\left(\left\{\left[\frac{i_s - 1}{T}, \frac{i_s}{T}\right]\right\}_{s=1}^{S}, \left\{\left[\frac{j_y - 1}{T}, \frac{j_y}{T}\right]\right\}_{y=1}^{C}\right), \text{ where} \tag{7}$$

$$\mathbf{C}\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right) = \max_{x_{s,y}} \sum_{s=1}^{S}\sum_{y=1}^{C} \Big(\overline{k_s}\,\overline{r_y}\,(E_{s,y} + C_{s,y})_+ + \underline{k_s}\,\underline{r_y}\,(E_{s,y} + C_{s,y})_-$$

$$+2\overline{k_s}\,\overline{r_y}\sqrt{x_{s,y}(1 - x_{s,y})}\sqrt{V_{s,y}} - \underline{k_s}\,\underline{r_y} x_{s,y}(C_{s,y})_+ - \overline{k_s}\,\overline{r_y} x_{s,y}(C_{s,y})_-\Big) \tag{8a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} \underline{k_s} \leq 1, \quad \sum_{s=1}^{S} \overline{k_s} \geq 1, \quad \sum_{y=1}^{C} \underline{r_y} \leq 1, \quad \sum_{y=1}^{C} \overline{r_y} \geq 1, \tag{8b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y}\overline{k_s}\,\overline{r_y} x_{s,y}} \geq 1 - \rho^2, \quad (1 - \bar{\gamma}_{s,y}^2)^2 \leq x_{s,y} \leq 1 \quad \forall s \in [S], y \in [C], \tag{8c}$$

*where $(\cdot)_+ = \max\{\cdot, 0\}$, $(\cdot)_- = \min\{\cdot, 0\}$; $E_{s,y} = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$, $V_{s,y} = \mathbb{V}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$, $p_{s,y} = \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y]$, $C_{s,y} = M - E_{s,y} - \frac{V_{s,y}}{M - E_{s,y}}$, and $\bar{\gamma}_{s,y}^2 = 1 - (1 + (M - E_{s,y})^2/V_{s,y})^{-\frac{1}{2}}$. Equation (7) only takes $\mathbf{C}$'s value when it is feasible, and each $\mathbf{C}$ queried by Equation (7) is a **convex optimization**.*

Implications. Thm. 3 provides a fairness certificate for Problem 1 under two assumptions: (1) The loss function is bounded (by M). This assumption holds for several typical losses such as 0-1 loss and JSD loss. (2) The distribution shift between training and test distribution within each subpopulation is bounded by $\bar{\gamma}_{s,y}$, where $\bar{\gamma}_{s,y}$ is determined by the model's statistics on $\mathcal{P}$. In practice, this additional distance bound assumption generally holds, since $\bar{\gamma}_{s,y} \gg \rho$ for common choices of ρ.

In Thm. 3, we exploit three types of statistics of $h_\theta(\cdot)$ on $\mathcal{P}$ to compute the fairness certificates: the expectation $E_{s,y} = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$, the variance $V_{s,y} = \mathbb{V}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$, and the density $p_{s,y} = \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y]$, all of which are at the subpopulation level and a high-confidence estimation of them based on finite samples are tractable (Section 3.4).

Using Thm. 3, after determining the region granularity T, we can provide a fairness certificate for Problem 1 by solving T^{SC} convex optimization problems, each of which has SC decision variables. Note that the computation cost is independent of h_θ, and therefore we can numerically compute the certificate for large DNN models used in practice. Specifically, when $S = 2$ (binary sensitive attribute) or $C = 2$ (binary classification) which is common in the fairness evaluation setting, we can construct the region for only one dimension k_1 or r_1, and use $1 - k_1$ or $1 - r_1$ for the other dimension. Thus, for the typical setting $S = 2, C = 2$, we only need to solve T^2 convex optimization problems.

Note that for Problem 2, our certificate in Thm. 2 is tight, whereas for Problem 1, our certificate in Thm. 3 is not. This is because in Problem 1, extra distribution shift exists within each subpopulation, i.e., $\Pr(X_o|Y, X_s)$ changes from $\mathcal{P}$ to $\mathcal{Q}$, and to bound such shift, we need to leverage Thm. 2.2 in [47] which has no tightness guarantee. Future work providing tighter bounds than [47] can be seamlessly incorporated into our framework to tighten our fairness certificate for Problem 1.

3.4 Dealing with Finite Sampling Error

In Section 3.2 and Section 3.3, we present Thm. 2 and Thm. 3 that provide computable fairness certificates for sensitive shifting and general shifting scenarios respectively. In these theorems, we need to know the quantities related to the training distribution and trained $\mathcal{P}$ and model $h_\theta(\cdot)$:

$$E_{s,y} = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)], V_{s,y} = \mathop{\mathbb{V}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)], p_{s,y} = \mathop{\Pr}_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y]. \quad (9)$$

Section 3.3 further requires $C_{s,y}$ and $\bar{\gamma}_{s,y}$ which are functions of $E_{s,y}$ and $V_{s,y}$. However, a practical challenge is that common training distributions do not have an analytical expression that allows us to precisely compute these quantities. Indeed, we only have access to a finite number of individually drawn samples, i.e., the training dataset, from $\mathcal{P}$. Thus, we will provide high-confidence bounds for $E_{s,y}$, $V_{s,y}$, and $p_{s,y}$ in Lemma D.1 (stated in Appendix D.1).

For Thm. 2, we can replace $E_{s,y}$ in the objective by the upper bounds of $E_{s,y}$ and replace the concrete quantities of $p_{s,y}$ by interval constraints and the unit constraint $\sum_s \sum_y p_{s,y} = 1$, which again yields a convex optimization that can be effectively solved. For Thm. 3, we compute the confidence intervals of $C_{s,y}$ and $\rho_{s,y}$, then plug in either the lower bounds or the upper bounds to the objective (8a) based on the coefficient, and finally replace the concrete quantities of $p_{s,y}$ by interval constraints and the unit constraint $\sum_s \sum_y p_{s,y} = 1$. The resulting optimization is proved to be convex and provides an upper bound for any possible values of $E_{s,y}$, $V_{s,y}$, and $p_{s,y}$ within the confidence intervals. We defer the statement of Thm. 2 and Thm. 3 considering finite sampling error to Appendix D.2. To this point, we have presented our framework for computing high-confidence fairness certificates given access to model $h_\theta(\cdot)$ and a finite number of samples drawn from $\mathcal{P}$.

4 Experiments

In this section, we evaluate the certified fairness under both *sensitive shifting* and *general shifting* scenarios on six real-world datasets. We observe that under the sensitive shifting, our certified fairness bound is *tight* (Section 4.1); while the bound is less tight under general shifting (Section 4.2) which depends on the tightness of generalization bounds within each subpopulation (details in Section 3.3). In addition, we show that our certification framework can flexibly integrate more constraints on $\mathcal{Q}$, leading to a tighter fairness certification (Section 4.3). Finally, we compare our certified fairness bound with existing distributional robustness bound [43] (section 4.4), since both consider a shifted distribution while our bound is optimized with an additional fairness constraint which is challenging to be directly integrated to the existing distributional robustness optimization. We show that with the fairness constraint and our optimization approach, our bound is much tighter.

Dataset & Model. We validate our certified fairness on *six* real-world datasets: Adult [3], Compas [2], Health [19], Lawschool [48], Crime [3], and German [3]. Details on the datasets and data processing steps are provided in Appendix E.1. Following the standard setup of fairness evaluation in the literature [39, 38, 31, 42], we consider the scenario that the sensitive attributes and labels take binary values. The ReLU network composed of 2 hidden layers of size 20 is used for all datasets.

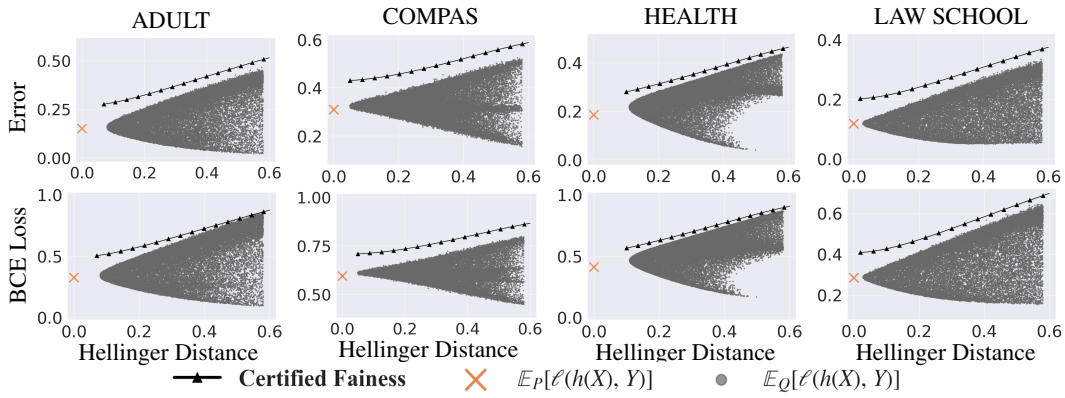


Figure 1: Certified fairness with sensitive shifting. Grey points are results on generated distributions ($\mathcal{Q}$) and the black line is our fairness certificate based on Thm. 2. We observe that our fairness certificate is usually tight.

Fairness Certification. We perform vanilla model training and then leverage our fairness certification framework to calculate the fairness certificate. Concretely, we input the trained model information on $\mathcal{P}$ and the framework would output the fairness certification for both sensitive shifting and general shifting scenarios following Thm. 2 and Thm. 3, respectively.

Code, model, and all experimental data are publicly available at `https://github.com/AI-secure/Certified-Fairness`.

4.1 Certified Fairness with Sensitive Shifting

Generating Fair Distributions. To evaluate how well our certificates capture the fairness risk in practice, we compare our certification bound with the empirical loss evaluated on randomly generated $30,000$ fairness constrained distributions $\mathcal{Q}$ shifted from $\mathcal{P}$. The detailed steps for generating fairness constrained distributions $\mathcal{Q}$ are provided in Appendix E.2. Under sensitive shifting, since each subpopulation divided by the sensitive attribute and label does not change (Section 2.1), we tune only the portion of each subpopulation $q_{s,y}$ satisfying the base rate fairness constraint, and then sample from each subpopulation of $\mathcal{P}$ individually according to the proportion $q_{s,y}$. In this way, our protocols can generate distributions with different combinations of subpopulation portions. If the classifier is biased toward one subpopulation (i.e., it achieves high accuracy in the group but low accuracy in others), the worst-case accuracy on generated distribution is low since the portion of the biased subpopulation in the generated distribution can be low; in contrast, a fair classifier which performs uniformly well for each group can achieve high worst-case accuracy (high certified fairness). Therefore, we believe that our protocols can demonstrate real-world training distribution bias as well as reflect the model's unfairness and certification tightness in real-world scenarios.

Results. We report the classification error (Error) and BCE loss as the evaluation metric. Figure 1 illustrates the certified fairness on Adult, Compas, Health, and Lawschool under sensitive shifting. More results on two relatively small datasets (Crime, German) are shown in Appendix E.5. From the results, we see that our certified fairness is tight in practice.

4.2 Certified Fairness with General Shifting

In the general shifting scenario, we similarly randomly generate $30,000$ fair distributions $\mathcal{Q}$ shifted from $\mathcal{P}$. Different from sensitive shifting, the distribution conditioned on sensitive attribute X_s and label Y can also change in this scenario. Therefore, we construct another distribution $\mathcal{Q}'$ disjoint with $\mathcal{P}$ on non-sensitive attributes and mix $\mathcal{P}$ and $\mathcal{Q}'$ in each subpopulation individually guided by mixing parameters satisfying fair base rate constraint. Detailed generation steps are given in Appendix E.2. Since the fairness certification for general shifting requires bounded loss, we select classification error (Error) and Jensen-Shannon loss (JSD Loss) as the evaluation metric. Figure 2 illustrates the certified fairness with classification error metric under general shifting. Results of JSD loss and more results on two relatively small datasets (Crime, German) are in Appendix E.5.

4.3 Certified Fairness with Additional Non-Skewness Constraints

In Section 2.1, we discussed that to represent different real-world scenarios we can add more constraints such as Equation (1) to prevent the skewness of $\mathcal{Q}$, which can be flexibly incorporated into our certificate framework. Concretely, for sensitive shifting, we only need to add one more

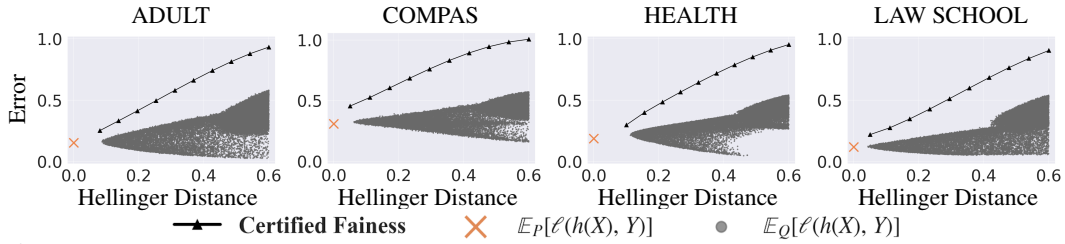
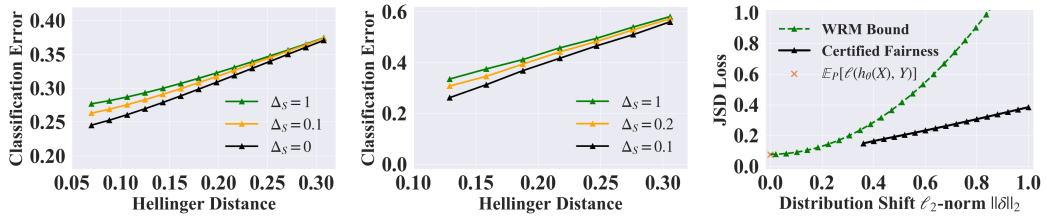


Figure 2: Certified fairness with general shifting. Grey points are results on generated distributions ($\mathcal{Q}$) and the black line is our fairness certificate based on Thm. 3. We observe that our fairness certificate is non-trivial.



(a) Sentitive Shifting (b) General Shifting (c) Comparison with WRM

Figure 3: Certified fairness with additional non-skewness constraints on Adult dataset is shown in (a) (b). Δ_s controls the skewness of $\mathcal{Q}$ ($|\Pr_{(X,Y)\sim\mathcal{Q}}[X_s = 0] - \Pr_{(X,Y)\sim\mathcal{Q}}[X_s = 1]| \leq \Delta_s$). More analysis in Section 4.3. In (c), we compare our certified fairness bound with the distributional robustness bound [43]. More analysis in Section 4.4.

box constraint[3] $0.5 - \Delta_s/2 \leq k_s \leq 0.5 + \Delta_s/2$ where Δ_s is a parameter controlling the skewness of $\mathcal{Q}$, which still guarantees convexity. For general shifting, we only need to modify the region partition step[3], where we split $[0.5 - \Delta_s/2, 0.5 + \Delta_s/2]$ instead of $[0, 1]$. The certification results with additional constraints are in Figures 3(a) and 3(b), which suggests that if the added constraints are strict (i.e., smaller Δ_s), the bound is tighter. More constraints w.r.t. labels can also be handled by our framework and the corresponding results as well as results on more datasets are in Appendix E.6.

4.4 Comparison with Distributional Robustness Bound

To the best of our knowledge, there is no existing work providing *certified fairness* on the end-to-end model performance. Thus, we try to compare our bound with the distributional robustness bound since both consider certain distribution shifts. However, it is challenging to directly integrate the fairness constraints into existing bounds. Therefore, we compare with the state-of-the-art distributional robustness certification WRM [43], which solves the similar optimization problem as ours except for the fairness constraint. For fair comparison, we construct a synthetic dataset following [43], on which there is a one-to-one correspondence between the Hellinger and Wasserstein distance used by WRM. We randomly select one dimension as the sensitive attribute. Since WRM has additional assumptions on smoothness of models and losses, we use JSD loss and a small ELU network with 2 hidden layers of size 4 and 2 following their setting. More implementation details are in Appendix E.4. Results in Figure 3(c) suggest that 1) our certified fairness bound is much tighter than WRM given the additional fairness distribution constraint and our optimization framework; 2) with additional fairness constraint, our certificate problem could be infeasible under very small distribution distances since the fairness constrained distribution $\mathcal{Q}$ does not exist near the skewed original distribution $\mathcal{P}$; 3) with the fairness constraint, we provide non-trivial fairness certification bound even when the distribution shift is large.

5 Conclusion

In this paper, we provide the first *fairness certification* on end-to-end model performance, based on a fairness constrained distribution which has bounded distribution distance from the training distribution. We show that our fairness certification has strong connections with existing fairness notions such as group parity, and we provide an effective framework to calculate the certification under different scenarios. We provide both theoretical and empirical analysis of our fairness certification.

Acknowledgements. MK, LL, and BL are partially supported by the NSF grant No.1910100, NSF CNS No.2046726, C3 AI, and the Alfred P. Sloan Foundation. YL is partially supported by the NSF grants IIS-2143895 and IIS-2040800.

[3]Note that such modification is only viable when sentive attributes take binary values, which is the typical scenario in the literature of fairness evaluation [39, 38, 31, 42].

References

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

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[3] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[4] Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair normalizing flows. *arXiv preprint arXiv:2106.05937*, 2021.

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

[6] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA):1–27, 2019.

[7] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[9] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *Advances in Neural Information Processing Systems*, 2022.

[10] YooJung Choi, Meihua Dang, and Guy Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. *arXiv preprint arXiv:2009.09031*, 2020.

[11] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[12] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

[13] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

[14] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[15] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

[16] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[17] Jeffrey A Gliner. Reviewing qualitative research: Proposed criteria for fairness and rigor. *The Occupational Therapy Journal of Research*, 14(2):78–92, 1994.

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

[19] Kaggle Inc. Heritage health prize kaggle. `https://www.kaggle.com/c/hhp`, 2022.

[20] Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. On transferability of bias mitigation effects in language model fine-tuning. *arXiv preprint arXiv:2010.12864*, 2020.

[21] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pages 749–758. PMLR, 2020.

[22] Thomas Kehrenberg, Myles Bartlett, Oliver Thomas, and Novi Quadrianto. Null-sampling for interpretable and fair representations. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[24] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[25] Henry Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[26] Jiachun Liao, Chong Huang, Peter Kairouz, and Lalitha Sankar. Learning generative adversarial representations (gap) under fairness and censoring constraints. *arXiv preprint arXiv:1910.00411*, 1, 2019.

[27] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

[28] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[30] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[31] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25773–25784. Curran Associates, Inc., 2021.

[32] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[33] Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 263–270, New York, NY, USA, 2019. Association for Computing Machinery.

[34] Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39, 2020.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[36] Momchil Peychev, Anian Ruoss, Mislav Balunović, Maximilian Baader, and Martin Vechev. Latent space smoothing for individually fair representations. *arXiv preprint arXiv:2111.13650*, 2021.

[37] Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *Advances in Neural Information Processing Systems*, 33:7584–7596, 2020.

[40] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pages 746–761. Springer, 2020.

[41] Shahar Segal, Yossi Adi, Benny Pinkas, Carsten Baum, Chaya Ganesh, and Joseph Keshet. Fairness in the eyes of the data: Certifying machine-learning models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 926–935, 2021.

[42] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24535–24544. Curran Associates, Inc., 2021.

[43] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[44] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

[45] Ton Steerneman. On the total variation and hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

[46] Caterina Urban, Maria Christakis, Valentin Wüstholz, and Fuyuan Zhang. Perfectly parallel fairness certification of neural networks. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):1–30, 2020.

[47] Maurice Weber, Linyi Li, Boxin Wang, Zhikuan Zhao, Bo Li, and Ce Zhang. Certifying out-of-domain generalization for blackbox functions. In *International Conference on Machine Learning*. PMLR, 2022.

[48] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. *Law School Admission Council*, 1998.

[49] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[50] Samuel Yeom and Matt Fredrikson. Individual fairness revisited: Transferring techniques from adversarial robustness. *arXiv preprint arXiv:2002.07738*, 2020.

[51] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[52] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

[53] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019.

[54] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We clearly specify the assumptions and application scopes, if exists, of all our results.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss broader impacts in Appendix A.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We include complete proofs for all results in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include them in Appendix E.3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Indeed, we rigorously compute the confidence intervals due to the finite sampling (see Section 3.4), and guarantee that all fairness certificates reported in the paper holds with probability $\geq 90\%$.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide them in Appendix E.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We only use public data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We only use public data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

Contents

A Broader Impact **16**

B Omitted Background **16**

B.1 Hellinger Distance 16

B.2 Thm. 2.2 in [47] 16

C Proofs of Main Results **17**

C.1 Proof of Proposition 1 17

C.2 Proof of Thm. 1 17

C.3 Proof of Thm. 2 18

C.4 Proof of Lemma 3.1 20

C.5 Proof of Thm. 3 21

D Omitted Theorem Statements and Proofs for Finite Sampling Error **24**

D.1 Finite Sampling Confidence Intervals 24

D.2 Fairness Certification Statements with Finite Sampling 25

D.3 Proofs of Fairness Certification with Finite Sampling 26

E Experiments **28**

E.1 Datasets 28

E.2 Fair Base Rate Distribution Generation Protocol 29

E.3 Implementation Details of Our Fairness Certification 30

E.4 Implementation Details of WRM 30

E.5 More Results of Certified Fairness with Sensitive Shifting and General shifting . . 31

E.6 More Results of Certified Fairness with Additional Non-Skewness Constraints . . . 33

E.7 Fair Classifier Achieves High Certified Fairness 36

A Broader Impact

This paper aims to calculate a *fairness certificate* under some distributional fairness constraints on the performance of an end-to-end ML model. We believe that the rigorous fairness certificates provided by our framework will significantly benefit and advance social fairness in the era of deep learning. Especially, such fairness certificate can be directly used to measure the fairness of an ML model regardless the target domain, which means that it will measure the unique property of the model itself with theoretical guarantees, and thus help people understand the risks of existing ML models. As a result, the ML community may develop ML training algorithms that explicitly reduce the fairness risks by regularizing on this fairness certificate.

A possible negative societal impact may stem from the misunderstanding or inaccurate interpretation of our fairness certificate. As a first step towards distributional fairness certification, we define the fairness through the lens of worst-case performance loss on a fairness constrained distribution. This fairness definition may not explicitly imply an absoluate fairness guarantee under some other criterion. For example, it does not imply that for any possible individual input, the ML model will give fair prediction. We tried our best in Section 2 to define the certification goal, and the practitioners may need to understand this goal well to avoid misinterpretation or misuse of our fairness certification.

B Omitted Background

We illustrate omitted background in this appendix.

B.1 Hellinger Distance

As illustrated in the beginning of Section 3, our framework uses Hellinger distance to bound the distributional distance. A formal definition of Hellinger distance is as below.

Definition 3 (Hellinger Distance)**.** Let $\mathcal{P}$ and $\mathcal{Q}$ be distributions on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ that are absolutely continuous with respect to a reference measure μ with $\mathcal{P}, \mathcal{Q} \ll \mu$. The Hellinger distance between $\mathcal{P}$ and $\mathcal{Q}$ is defined as

$$H(\mathcal{P}, \mathcal{Q}) := \sqrt{\frac{1}{2} \int_{\mathcal{Z}} \left(\sqrt{p(z)} - \sqrt{q(z)} \right)^2 \mathrm{d}\mu(z)} \tag{10}$$

where $p = \frac{d\mathcal{P}}{d\mu}$ and $q = \frac{d\mathcal{Q}}{d\mu}$ are the Radon-Nikodym derivatives of $\mathcal{P}$ and $\mathcal{Q}$ with respect to μ, respectively. The Hellinger distance is independent of the choice of the reference measure μ.

Representative properties for the Hellinger distance are discussed in Section 3.

B.2 Thm. 2.2 in [47]

As mentioned in Section 3.3, we leverage Thm. 2.2 from [47] to upper bound the expected loss of $h_\theta(\cdot)$ in each shifted subpopulation $\mathcal{Q}_{s,y}$. Here we restate Thm. 2.2 for completeness.

Theorem 4 (Thm. 2.2, [47])**.** *Let $\mathcal{P}'$ and $\mathcal{Q}'$ denote two distributions supported on $\mathcal{X} \times \mathcal{Y}$, suppose that $0 \leq \ell(h_\theta(X), Y) \leq M$, then*

$$\begin{aligned} &\max_{\mathcal{Q}', \theta} \mathbb{E}_{(X,Y)\sim\mathcal{Q}'}[\ell(h_\theta(X), Y)] \quad \text{s.t.} \quad H(\mathcal{P}', \mathcal{Q}') \leq \rho \\ \leq &\mathbb{E}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)] + 2C_\rho \sqrt{\mathbb{V}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)]} + \\ &\rho^2(2-\rho^2) \left(M - \mathbb{E}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)] - \frac{\mathbb{V}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)]}{M - \mathbb{E}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)]} \right), \end{aligned} \tag{11}$$

where $C_\rho = \sqrt{\rho^2(1-\rho^2)^2(2-\rho^2)}$, for any given distance bound $\rho > 0$ that satisfies

$$\rho^2 \leq 1 - \left(1 + \frac{(M - \mathbb{E}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)])^2}{\mathbb{V}_{(X,Y)\sim\mathcal{P}'}[\ell(h_\theta(X), Y)]} \right)^{-1/2}. \tag{12}$$

This theorem provides a closed-form expression that upper bounds the mean loss of $h_\theta(\cdot)$ on shifted distribution (namely $\mathbb{E}_{\mathcal{Q}'}[\ell(h_\theta(X), Y)]$), given bounded Hellinger distance $H(\mathcal{P}, \mathcal{Q})$ and the mean E and variance V of loss on $\mathcal{P}$ under two mild conditions: (1) the function is positive and bounded (denote the upper bound by M); and (2) the distance $H(\mathcal{P}, \mathcal{Q})$ is not too large (specifically, $H(\mathcal{P}, \mathcal{Q})^2 \leq \bar{\gamma}^2 := 1 - (1 + (M - E)^2/V)^{-\frac{1}{2}}$). Since Thm. 4 holds for arbitrary models and loss functions $\ell(h_\theta(\cdot), \cdot)$ as long as the function value is bounded by $[0, M]$, using Thm. 4 allows us to provide a generic and succinct fairness certificate in Thm. 3 for general shifting case that holds for generic models including DNNs without engaging complex model architectures. Indeed, we only need to query the mean and variance under $\mathcal{P}$ for the given model to compute the certificate in Thm. 4, and this benefit is also inherited by our certification framework expressed by Thm. 3. Note that there is no tightness guarantee for this bound yet, which is also inherited by our Thm. 3.

C Proofs of Main Results

This appendix entails the complete proofs for Proposition 1, Thm. 1, Thm. 2, Lemma 3.1, and Thm. 3 in the main text. For complex proofs such as that for Thm. 3, we also provide high-level illustration before going into the formal proof.

C.1 Proof of Proposition 1

Proof of Proposition 1. Since each term $\Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) \neq Y | Y = y, X_s = i]$ is within $[0, \epsilon]$, we consider two cases: $y \neq 1$ and $y = 1$. If $y \neq 1$, $\Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | Y = y, X_s = i] \leq \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) \neq Y | Y = y, X_s = i] \leq \epsilon$ and so will be their differences for $X_s = i$ and $X_s = j$. If $y = 1$, $\Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | Y = y, X_s = i] = 1 - \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) \neq Y | Y = y, X_s = i] \in [1 - \epsilon, 1]$, and also the differences for $X_s = i$ and $X_s = j$ are always within ϵ. This proves ϵ-EO.

Now consider DP. We notice that for any a,

$$\Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | X_s = a] = \sum_{y=1}^{C} \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | Y = y, X_s = a] \cdot \Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = a]. \tag{13}$$

Thus,

$$\begin{aligned}
&\Big| \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | X_s = i] - \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | X_s = j] \Big| \\
\overset{(*)}{\leq} &\sum_{y=1}^{C} \Big| \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | Y = y, X_s = i] - \Pr_{(X,Y)\sim\mathcal{Q}}[h_\theta(X) = 1 | Y = y, X_s = j] \Big| \\
&\qquad \cdot \Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = i] \\
\leq &\sum_{y=1}^{C} \epsilon \Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = i] = \epsilon
\end{aligned}$$

which proves ϵ-DP, where $(*)$ leverages the fair base rate property of $\mathcal{Q}$ which gives $\Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = i] = \Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = j]$.

□

C.2 Proof of Thm. 1

Proof of Thm. 1. We first prove the key eq. (3).

$$\begin{aligned}
H(\mathcal{P}, \mathcal{Q}) \leq \rho &\iff H^2(\mathcal{P}, \mathcal{Q}) \leq \rho^2 \\
&\iff \frac{1}{2} \int_{\mathcal{Z}} \left(\sqrt{p(z)} - \sqrt{q(z)} \right)^2 \mathrm{d}\mu(z) \leq \rho^2 \\
&\iff \frac{1}{2} \left(\int_{\mathcal{Z}} p(z) \mathrm{d}\mu(z) + \int_{\mathcal{Z}} q(z) \mathrm{d}\mu(z) \right) - \int_{\mathcal{Z}} \sqrt{p(z)q(z)} \mathrm{d}\mu(z) \leq \rho^2
\end{aligned}$$

$$
\begin{aligned}
&\iff \int_{\mathcal{Z}} \sqrt{p(z)q(z)}\mathrm{d}\mu(z) \geq 1-\rho^2 \\
&\iff \sum_{i=1}^{N} \int_{\mathcal{Z}_i} \sqrt{p_i q_i} \cdot \sqrt{p_i(z)q_i(z)}\mathrm{d}\mu(z) \geq 1-\rho^2 \\
&\iff \sum_{i=1}^{N} \sqrt{p_i q_i}\left(1-H^2(\mathcal{P}_i,\mathcal{Q}_i)\right) \geq 1-\rho^2 \qquad (14)
\end{aligned}
$$

where $p_i(\cdot)$ and $q_i(\cdot)$ are density functions of subpopulation distributions $\mathcal{P}_i$ and $\mathcal{Q}_i$ respectively.

Then, we show that any feasible solution of Equation (2) satisfies the constraints in Equation (4). We let $\mathcal{Q}^\star$ and $\theta^\star$ denote a feasible solution of Equation (2), i.e.,

$$
H(\mathcal{P},\mathcal{Q}^\star) \leq \rho, \quad e_j(\mathcal{P},h_{\theta^\star}) \leq v_j \,\forall j \in [L], \quad g_j(\mathcal{Q}^\star) \leq u_j \,\forall j \in [M]. \qquad (15)
$$

We let $\{q_i^\star\}_{i=1}^N$ denote the proportions of $\mathcal{Q}^\star$ within each support partition $\mathcal{Z}_i$, and $\{\mathcal{Q}_i^\star\}_{i=1}^N$ the $\mathcal{Q}^\star$ in each subpopulation. By Equation (14), we have $1-\rho^2-\sum_{i=1}^N \sqrt{p_i q_i^\star}(1-\rho_i^2) \leq 0$ where $\rho_i = H^2(\mathcal{P}_i,\mathcal{Q}_i^\star)$. Note that by definition, $\sum_{i=1}^N q_i^\star = 1$ and $\forall i \in [N], q_i^\star \geq 0, \rho_i \geq 0$. Furthermore, by the implication relations stated in Thm. 1, for any $j \in [L]$, $e_j'(\{\mathcal{P}_i\}_{i=1}^N, \{p_i\}_{i=1}^N, h_{\theta^\star}) \leq v_j'$; and for any $j \in [M]$, $g_j'(\{\mathcal{Q}_i^\star\}_{i=1}^N, \{q_i^\star\}_{i=1}^N) \leq u_j'$. To this point, we have shown $\mathcal{Q}^\star$ and $\theta^\star$ satisfy all constraints in Equation (4), i.e., $\mathcal{Q}^\star$ and $\theta^\star$ is a feasible solution of Equation (4). Since Equation (4) expresses the optimal (maximum) solution, Equation (4) (in Thm. 1) $\geq$ Equation (2). □

C.3 Proof of Thm. 2

Proof of Thm. 2. The proof of Thm. 2 is composed of three parts: (1) the optimization problem provides a fairness certificate for Problem 2; (2) the certificate is tight; and (3) the optimization problem is convex.

(1) Suppose the maximum of Problem 2 is attained with the test distribution $\mathcal{Q}^\star$ in the sensitive shifting setting, then we decompose both $\mathcal{P}$ and $\mathcal{Q}^\star$ according to both the sensitive attribute and the label:

$$
\mathcal{P} = \sum_{s=1}^{S}\sum_{y=1}^{C} p_{s,y}\mathcal{P}_{s,y}, \quad \mathcal{Q}^\star = \sum_{s=1}^{S}\sum_{y=1}^{C} q_{s,y}^\star \mathcal{Q}_{s,y}^\star. \qquad (16)
$$

Since $\mathcal{Q}^\star$ is a fair base rate distribution, for any $i,j \in [S]$, $b_{i,y}^{\mathcal{Q}^\star} = b_{j,y}^{\mathcal{Q}^\star}$ where $b_{s,y}^{\mathcal{Q}^\star} = \Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y|X_s=s]$. As a result, $\Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y|X_s=s] = \Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y]$. Now we define

$$
k_s^\star := \Pr_{(X,Y)\sim\mathcal{Q}^\star}[X_s=s], \quad r_y^\star := \Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y], \qquad (17)
$$

and then

$$
q_{s,y}^\star = \Pr_{(X,Y)\sim\mathcal{Q}^\star}[X_s=s, Y=y] = \Pr_{(X,Y)\sim\mathcal{Q}^\star}[X_s=s] \cdot \Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y|X_s=s] = k_s^\star r_y^\star. \qquad (18)
$$

By the distance constraint in Problem 2 (namely $H(\mathcal{P},\mathcal{Q}^\star) \leq \rho$) and Equation (14), we have

$$
\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y}q_{s,y}^\star}\left(1-H^2(\mathcal{P}_{s,y},\mathcal{Q}_{s,y}^\star)\right) \geq 1-\rho^2. \qquad (19)
$$

Since there is only sensitive shifting, $H^2(\mathcal{P}_{s,y},\mathcal{Q}_{s,y}^\star) = 0$, given Equation (18), we have

$$
\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y}k_s^\star r_y^\star} \geq 1-\rho^2. \qquad (20)
$$

Now, we can observe that the $k_s^\star$ and $r_y^\star$ induced by $\mathcal{Q}^\star$ satisfy all constraints of Problem 2. For the objective,

Objective in Thm. 2

$$
\begin{aligned}
&=\sum_{s=1}^{S}\sum_{y=1}^{C} k_s^\star r_s^\star \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X),Y)]\\
&=\sum_{s=1}^{S}\sum_{y=1}^{C} q_{s,y}^\star \mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}^\star}[\ell(h_\theta(X),Y)] \qquad \text{(by Equation (18) and } H^2(\mathcal{P}_{s,y},\mathcal{Q}_{s,y}^\star)=0)\\
&=\mathbb{E}_{(X,Y)\sim\mathcal{Q}^\star}[\ell(h_\theta(X),Y)]\\
&=\text{Optimal value of Problem 2.}
\end{aligned}
$$

Therefore, the *optimal* value of Thm. 2 will be larger or equal to the optimal value of Problem 2 which concludes the proof of the first part.

(2) Suppose the optimal value of Thm. 2 is attained with $k_s^\star$ and $r_y^\star$. We then construct $\mathcal{Q}^\star=\sum_{s=1}^{S}\sum_{y=1}^{C} k_s^\star r_y^\star \mathcal{P}_{s,y}$. We now inspect each constraint of Problem 2. The constraint $\mathrm{dist}(\mathcal{P},\mathcal{Q}^\star)\le\rho$ is satisfied because $1-\rho^2-\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s^\star r_y^\star}\le 0$ is satisfied as a constraint of Thm. 2. Apparently, $\mathcal{P}_{s,y}=\mathcal{Q}_{s,y}^\star$. Then, $\mathcal{Q}^\star$ is a fair base rate distribution because

$$
b_{s,y}^{\mathcal{Q}^\star}=\Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y=y|X_s=s]=\frac{k_s^\star r_y^\star}{k_s^\star}=r_y^\star \tag{21}
$$

is a constant across all $s\in[S]$. Thus, $\mathcal{Q}^\star$ satisfies all constraints of Problem 2 and

$$
\begin{aligned}
&\text{Optimal objective of Problem 2}\\
\ge&\mathbb{E}_{(X,Y)\sim\mathcal{Q}^\star}[\ell(h_\theta(X),Y)]\\
=&\sum_{s=1}^{S}\sum_{y=1}^{C} k_s^\star r_y^\star \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X),Y)]\\
=&\sum_{s=1}^{S}\sum_{y=1}^{C} k_s^\star r_y^\star E_{s,y}=\text{Optimal objective of Thm. 2.}
\end{aligned} \tag{22}
$$

Combining with the conclusion of the first part, we know optimal values of Thm. 2 and Problem 2 match, i.e., the certificate is tight.

(3) Inspecting the problem definition in Thm. 2, we find the objective and all constraints but the last one are linear. Therefore, to prove the convexity of the optimization problem, we only need to show that the last constraint

$$
1-\rho^2-\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s r_y}\le 0 \tag{23}
$$

is convex with respect to k_s and r_y. Given two arbitrary feasible pairs of k_s and r_y satisfying Equation (23), namely (k_s^a,r_y^a) and (k_s^b,r_y^b), we only need to show that (k_s^m,r_y^m) also satisfies Equation (23), where $k_s^m=(k_s^a+k_s^b)/2$, $r_y^m=(r_y^a+r_y^b)/2$. Indeed,

$$
\begin{aligned}
&1-\rho^2-\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s^m r_y^m}\\
=&1-\rho^2-\frac{1}{2}\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}}\cdot\sqrt{k_s^a+k_s^b}\cdot\sqrt{r_y^a+r_y^b}\\
\le&1-\rho^2-\frac{1}{2}\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}}\cdot\left(\sqrt{k_s^a r_y^a}+\sqrt{k_s^b r_y^b}\right) \qquad \text{(Cauchy's inequality)}\\
=&\frac{1}{2}\left(1-\rho^2\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s^a r_y^a}\right)+\frac{1}{2}\left(1-\rho^2\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s^b r_y^b}\right)\\
\le&0.
\end{aligned}
$$

□

C.4 Proof of Lemma 3.1

Proof of Lemma 3.1. The proof of Lemma 3.1 is composed of two parts: (1) the optimization problem provides a fairness certificate for Problem 1; and (2) the certificate is tight. The high-level proof sketch is similar to the proof of Thm. 2.

(1) Suppose that the maximum of Problem 1 is attained with the test distribution $\mathcal{Q}^\star$ under the general shifting setting, then we decompose both $\mathcal{P}$ and $\mathcal{Q}^\star$ according to both the sensitive attribute and the label:

$$\mathcal{P} = \sum_{s=1}^{S}\sum_{y=1}^{C} p_{s,y}\mathcal{P}_{s,y}, \quad \mathcal{Q}^\star = \sum_{s=1}^{S}\sum_{y=1}^{C} q^\star_{s,y}\mathcal{Q}^\star_{s,y}. \tag{24}$$

Unlike sensitive shifting setting, in general shifting setting, here the subpopulation of $\mathcal{Q}^\star$ is $\mathcal{Q}^\star_{s,y}$ instead of $\mathcal{P}_{s,y}$ due to the existence of distribution shifting within each subpopulation.

Following the same argument as in the first part proof of Thm. 2, since $\mathcal{Q}^\star$ is a fair base rate distribution, we can define

$$k^\star_s := \Pr_{(X,Y)\sim\mathcal{Q}^\star}[X_s = s], \quad r^\star_y := \Pr_{(X,Y)\sim\mathcal{Q}^\star}[Y = y], \tag{25}$$

and write

$$\mathcal{Q}^\star := \sum_{s=1}^{S}\sum_{y=1}^{C} k^\star_s r^\star_y \mathcal{Q}^\star_{s,y} \tag{26}$$

since $q^\star_{s,y} = k^\star_s r^\star_y$. We also define $\rho^\star_{s,y} = H(\mathcal{P}_{s,y}, \mathcal{Q}^\star_{s,y})$. Now we show these $k^\star_s, r^\star_y, \mathcal{Q}^\star_{s,y}, \rho^\star_{s,y}$ along with model parameter θ constitute a feasible point of Equation (6), and the objectives of Equation (6) and Problem 2 are the same given $\mathcal{Q}^\star$.

- (Feasibility)
  There are three constraints in Equation (6). By the definition of $k^\star_s$ and $r^\star_y$, naturally Equation (6b) is satisfied. Then, according to Equation (14) and the definifition of $\rho^\star_{s,y}$ above, Equation (6c) and Equation (6d) are satisfied.
- (Objective Equality)

$$\begin{aligned}\text{Equation (6a)} &= \sum_{s=1}^{S}\sum_{y=1}^{C} k^\star_s r^\star_y \mathbb{E}_{(X,Y)\sim\mathcal{Q}^\star_{s,y}}[\ell(h_\theta(X), Y)] \\ &= \sum_{s=1}^{S}\sum_{y=1}^{C} q^\star_{s,y} \mathbb{E}_{(X,Y)\sim\mathcal{Q}^\star_{s,y}}[\ell(h_\theta(X), Y)] \\ &= \mathbb{E}_{(X,Y)\sim\mathcal{Q}^\star}[\ell(h_\theta(X), Y)] = \text{Optimal value of Problem 1.}\end{aligned} \tag{27}$$

As a result, the optimal value of Equation (6) is larger than or equal to the optimal value of Problem 1, and hence the optimization problem encoded by Equation (6) provides a fairness certificate.

(2) To prove the tightness of the certificate, we only need to show that the optimal value of the optimization problem in Equation (6) is also attainable by the original Problem 1.

Suppose that the optimal objective of Equation (6) is achieved by optimizable parameters $k^\star_s, r^\star_y, \mathcal{Q}^\star$, and $\rho^\star_{s,y}$. Then, we construct $\mathcal{Q}^\dagger = \sum_{s=1}^{S}\sum_{y=1}^{C} k^\star_s r^\star_y \mathcal{Q}^\star_{s,y}$. We first show that $\mathcal{Q}^\dagger$ is a feasible point of Problem 1, and then show that the objective given $\mathcal{Q}^\dagger$ is equal to the optimal objective of Equation (6).

- (Feasibility)
  There are two constraints in Problem 1: the bounded distance constraint and the fair base rate constraint. The bounded distance constraint is satisfied due to applying Equation (14) along with Equations (6c) and (6d). The fair base rate constraint is satisfied following the same deduction as in Equation (21).

- (Objective Equality)

$$\text{Objective Problem 1} = \mathbb{E}_{(X,Y)\sim\mathcal{Q}^\dagger}[\ell(h_\theta(X),Y)] = \sum_{s=1}^{S}\sum_{y=1}^{C} k_s^\star r_y^\star \mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}^\star}[\ell(h_\theta(X),Y)]$$
$$= \text{Optimal value of Equation (6).}$$

Thus, the optimal value of the optimization problem in Equation (6) is attainable also by the original Problem 1 which concludes the tightness proof.

□

C.5 Proof of Thm. 3

High-Level Illustration. The starting point of our proof is Lemma 3.1, where we have shown a fairness certificate for Problem 1 (general shifting setting). Then, we plug in Thm. 2.2 in [47] (stated as Thm. 4 in Appendix B.2) to upper bound the expected loss within each sub-population. Now, we get an optimization problem involving k_s, r_y, and $\rho_{s,y}$ that upper bounds the optimization problem in Lemma 3.1. In this optimization problem, we find k_s and r_y are bounded in $[0,1]$, and once these two variables are fixed, the optimization with respect to $x_{s,y} := (1-\rho_{s,y}^2)^2$ becomes convex. Using this observation, we propose to partition the feasible space of k_s and r_y into sub-regions and solve the convex optimization within each region bearing some degree relaxation, which yields Thm. 3.

Proof of Thm. 3. The proof is done stage-wise: starting from Lemma 3.1, we apply relaxation and derive a subsequent optimization problem that upper bounds the previous one stage by stage, until we get the final expression in Thm. 3.

To demonstrate the proof, we first define the optimization problems at each stage, then prove the relaxations between each adjacent stage, and finally show that the last optimization problem contains a finite number of **C**'s values where each **C** is a convex optimization, so that the final optimization problem provides a computable fairness certificate.

We define these quantities, for $s \in [S], y \in [C]$:

$$\begin{aligned}
&E_{s,y} = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X),Y)], \quad V_{s,y} = \mathbb{V}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X),Y)],\\
&p_{s,y} = \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y], \quad C_{s,y} = M - E_{s,y} - \frac{V_{s,y}}{M - E_{s,y}},\\
&\bar{\gamma}_{s,y}^2 = 1 - (1 + (M - E_{s,y})^2/V_{s,y})^{-\frac{1}{2}}.
\end{aligned} \tag{28}$$

Given $\rho > 0$ and the above quantities, the optimization problem definitions are:

- Lemma 3.1:

$$\max_{k_s, r_y, \mathcal{Q}, \rho_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X),Y)] \tag{29a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C], \tag{29b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y}(1-\rho_{s,y}^2) \geq 1 - \rho^2 \tag{29c}$$

$$H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \rho_{s,y} \quad \forall s \in [S], y \in [C]. \tag{29d}$$

- After applying Thm. 4:

$$\max_{k_s, r_y, \rho_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \left(E_{s,y} + 2\sqrt{\rho_{s,y}^2(1-\rho_{s,y}^2)^2(2-\rho_{s,y}^2)}\sqrt{V_{s,y}} + \rho_{s,y}^2(2-\rho_{s,y}^2)C_{s,y}\right) \tag{30a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C], \tag{30b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y}(1-\rho_{s,y}^2) \geq 1-\rho^2, \tag{30c}$$

$$0 \leq \rho_{s,y} \leq \bar{\gamma}_{s,y}. \tag{30d}$$

- After variable transform $x_{s,y} := (1-\rho_{s,y}^2)^2$:

$$\max_{k_s, r_y, x_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \left(E_{s,y} + 2\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}} + (1-x_{s,y})C_{s,y} \right) \tag{31a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C], \tag{31b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y x_{s,y}} \geq 1-\rho^2, \tag{31c}$$

$$(1-\bar{\gamma}_{s,y}^2)^2 \leq x_{s,y} \leq 1 \quad \forall s \in [S], y \in [C]. \tag{31d}$$

- After feasible region partitioning on k_s and r_y:

$$\max_{\{i_s \in [T]: s \in [S]\}, \{j_y \in [T]: y \in [C]\}} \mathsf{C}' \left(\left\{ \left[\frac{i_s - 1}{T}, \frac{i_s}{T} \right] \right\}_{s=1}^{S}, \left\{ \left[\frac{j_y - 1}{T}, \frac{j_y}{T} \right] \right\}_{y=1}^{C} \right), \text{ where} \tag{32a}$$

$$\mathsf{C}' \left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C} \right) = \tag{32b}$$

$$\max_{\underline{k_s} \leq k_s \leq \overline{k_s}, \underline{r_y} \leq r_y \leq \overline{r_y}, x_{s,y}} \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \left(E_{s,y} + 2\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}} + (1-x_{s,y})C_{s,y} \right)$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \tag{32c}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y x_{s,y}} \geq 1-\rho^2, \tag{32d}$$

$$(1-\bar{\gamma}_{s,y}^2)^2 \leq x_{s,y} \leq 1 \quad \forall s \in [S], y \in [C]. \tag{32e}$$

- Final quantity in Thm. 3:

$$\max_{\{i_s \in [T]: s \in [S]\}, \{j_y \in [T]: y \in [C]\}} \mathsf{C} \left(\left\{ \left[\frac{i_s - 1}{T}, \frac{i_s}{T} \right] \right\}_{s=1}^{S}, \left\{ \left[\frac{j_y - 1}{T}, \frac{j_y}{T} \right] \right\}_{y=1}^{C} \right), \text{ where} \tag{33a}$$

$$\mathsf{C} \left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C} \right) = \max_{x_{s,y}} \sum_{s=1}^{S}\sum_{y=1}^{C} \Big(\overline{k_s}\,\overline{r_y} \left(E_{s,y} + C_{s,y}\right)_+ + \tag{33b}$$

$$\underline{k_s}\,\underline{r_y} \left(E_{s,y} + C_{s,y}\right)_- + 2\overline{k_s}\,\overline{r_y}\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}} - \underline{k_s}\,\underline{r_y} x_{s,y}(C_{s,y})_+ - \overline{k_s}\,\overline{r_y} x_{s,y}(C_{s,y})_- \Big)$$

$$\text{s.t.} \quad \sum_{s=1}^{S} \underline{k_s} \leq 1, \quad \sum_{s=1}^{S} \overline{k_s} \geq 1, \quad \sum_{y=1}^{C} \underline{r_y} \leq 1, \quad \sum_{y=1}^{C} \overline{r_y} \geq 1, \tag{33c}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} \overline{k_s}\,\overline{r_y} x_{s,y}} \geq 1-\rho^2, \tag{33d}$$

$$(1-\bar{\gamma}_{s,y}^2)^2 \leq x_{s,y} \leq 1 \quad \forall s \in [S], y \in [C]. \tag{33e}$$

We have this relation:

$$\text{Problem 1} \underbrace{\leq}_{\text{Lemma 3.1}} (29) \underbrace{\overset{\substack{\text{(when } \ell(h_\theta(X),Y) \in [0,M] \\ \text{and } H(\mathcal{P}_{s,y},\mathcal{Q}_{s,y}) \leq \bar{\gamma}_{s,y})}}{\leq}}_{\textbf{(A)}} (30) \underbrace{=}_{\textbf{(B)}} (31) \underbrace{=}_{\textbf{(C)}} (32) \underbrace{\leq}_{\textbf{(D)}} (33). \tag{34}$$

Thus, when $H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \bar{\gamma}_{s,y}$ and $\sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \ell(h_\theta(X), Y) \leq M$, given that ℓ is a non-negative loss by Section 2, we can see Equation (33), i.e., the expression in Thm. 3's statement, upper bounds Problem 1, i.e., provides a fairness certificate for Problem 1. The proofs of these equalities/inequalities are in the following parts labeled by **(A)**, **(B)**, **(C)**, and **(D)** respectively.

Now we show that each **C** queried by Equation (7) (or equally Equation (33a)) is a convex optimization. Inspecting **C**'s objective, with respect to the optimizable variable $x_{s,y}$, we find that the only non-linear term in the objective is $\sum_{s=1}^{S}\sum_{y=1}^{C} 2\overline{k_s}\overline{r_y}\sqrt{V_{s,y}}\sqrt{x_{s,y}(1-x_{s,y})}$. Consider the function $f(x) = \sqrt{x(1-x)}$. Define $g(y) = \sqrt{y}$ and $h(x) = x(1-x)$, and then $f(x) = g(h(x))$. Thus, $f'(x) = g'(h(x))h'(x)$ and $f''(x) = g''(h(x))h'(x)^2 + g'(h(x))h''(x)$. Notice that $g''(h(x)) \leq 0$, $g'(h(x)) > 0$, and $h''(x) < 0$ for $x \in (0,1]$. Thus, $f''(x) \leq 0$. Since f is twice differentiable in $(0,1]$, we can conclude that f is concave and so does the objective of Equation (7). Inspecting **C**'s constraints, we observe that the only non-linear constraint is $\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}} \geq 1-\rho^2$. Due to the concavity of function $x \mapsto \sqrt{x}$, we have $\sqrt{p_{s,y}\overline{k_s}\overline{r_y}(x_{s,y}^a + x_{s,y}^b)/2} \geq \frac{1}{2}\left(\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}^a} + \sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}^b}\right)$ for any two feasible points $x_{s,y}^a$ and $x_{s,y}^b$. Thus, this non-linear constraint defines a convex region. To this point, we have shown that **C**'s objective is concave and **C**'s constraints are convex, given that **C** is a maximization problem, **C** is a convex optimization. □

Under the assumptions that $\ell(h_\theta(X), Y) \in [0, M]$ and $H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \bar{\gamma}_{s,y}$:

(A) *Proof of Equation* (29) $\leq$ *Equation* (30).
Given Equation (29d), for each $\mathcal{Q}_{s,y}$, applying Thm. 4, we get

$$\mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X),Y)] \leq E_{s,y} + 2\sqrt{\rho_{s,y}^2(1-\rho_{s,y}^2)^2(2-\rho_{s,y}^2)}\sqrt{V_{s,y}} + \rho_{s,y}^2(2-\rho_{s,y}^2)C_{s,y}. \tag{35}$$

Plugging this inequality into all $\mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X),Y)]$ in Equation (29a), we obtain Equation (30). □

(B) *Proof of Equation* (30) = *Equation* (31).
By Equation (30d), $\rho_{s,y} \in [0,1]$. Therefore, $x_{s,y} := (1-\rho_{s,y}^2)^2$ is a one-to-one mapping, and we can use $x_{s,y}$ to parameterize $\rho_{s,y}$, which yields Equation (31). □

(C) *Proof of Equation* (31) = *Equation* (32).
From Equation (31b), we notice that the feasible range of k_s and r_y is subsumed by $[0,1]$. We now partition this region $[0,1]$ for each variable to T sub-regions: $[(i-1)/T, i/T]$, $i \in [T]$, and then consider the maximum value across all the combinations of each sub-region for variables k_s and r_y, when feasible. As a result, Equation (31) can be written as the maximum over all such sub-problems where k_s's and r_y's enumerate all possible sub-region combinations, which is exactly encoded by Equation (32). □

(D) *Proof of Equation* (32) $\leq$ *Equation* (33).
We only need to show that when $\mathbf{C}'\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right)$ is feasible,

$$\mathbf{C}'\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right) \leq \mathbf{C}\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right). \tag{36}$$

Since both $\mathbf{C}'$ and **C** are maximization problem, we only need to show that the objective of **C** upper bounds that of $\mathbf{C}'$, and the constraints of $\mathbf{C}'$ are equal or relaxations of those of **C**.

For the objective, given that $\underline{k_s} \le k_s \le \overline{k_s}$ and $\underline{r_y} \le r_y \le \overline{r_y}$, for any $x_{s,y}$, We observe that

$$\begin{aligned} k_s r_y (E_{s,y} + C_{s,y}) &\le \overline{k_s}\,\overline{r_y}\,(E_{s,y} + C_{s,y})_+ + \underline{k_s}\,\underline{r_y}\,(E_{s,y} + C_{s,y})_-\,, \\ k_s r_y \cdot \left(2\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}}\right) &\le \overline{k_s}\,\overline{r_y} \cdot \left(2\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}}\right), \\ -k_s r_y C_{s,y} x_{s,y} &\le -\underline{k_s}\,\underline{r_y} x_{s,y}(C_{s,y})_+ - \overline{k_s}\,\overline{r_y} x_{s,y}(C_{s,y})_-, \end{aligned} \tag{37}$$

and by summing up all these terms for all $s \in [S]$ and $y \in [C]$, the LHS would be the objective of $\mathbf{C}'$ and the RHS would be the objective of $\mathbf{C}$. Hence, $\mathbf{C}$'s objective upper bounds that of $\mathbf{C}'$. For the constraints, similarly, given that $\underline{k_s} \le k_s \le \overline{k_s}$ and $\underline{r_y} \le r_y \le \overline{r_y}$, we have

$$\text{(32c)} \sum_{s=1}^{S} k_s = 1,\ \sum_{y=1}^{C} r_y = 1 \Longrightarrow \sum_{s=1}^{S} \underline{k_s} \le 1,\ \sum_{s=1}^{S} \overline{k_s} \ge 1,\ \sum_{y=1}^{C} \underline{r_y} \le 1,\ \sum_{y=1}^{C} \overline{r_y} \ge 1 \text{(33c)},$$

$$\text{(32d)} \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y x_{s,y}} \ge 1 - \rho^2 \Longrightarrow \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y}\overline{k_s}\,\overline{r_y} x_{s,y}} \ge 1 - \rho^2 \text{(33d)},$$

(32e) is as same as (33e),

which implies that all feasible solutions of $\mathbf{C}'$ are also feasible for $\mathbf{C}$. Combining with the fact that for any solution of $\mathbf{C}'$, its objective value $\mathbf{C}$ is greater than or equal to that of $\mathbf{C}'$ as shown above, we have Equation (36) which concludes the proof. □

D Omitted Theorem Statements and Proofs for Finite Sampling Error

D.1 Finite Sampling Confidence Intervals

Lemma D.1. *Let $\hat{P}$ be set of i.i.d. finite samples from $\mathcal{P}$, and let $\hat{P}_{s,y} := \{(X_i, Y_i) \in \hat{P} : (X_i)_s = s, Y_i = y\}$ for any $s \in [S], y \in [C]$. Let $\ell : (\hat{y}, y) \to [0, M]$ be a loss function. We define $\hat{L}_n = \frac{1}{|\hat{P}_{s,y}|}\sum_{(X_i,Y_i)\in\hat{P}_{s,y}} \ell(h_\theta(X_i), Y_i)$, $s_n^2 = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}^{n} \left(\ell(h_\theta(X_i), Y) - \ell(h_\theta(X_j), Y)\right)^2$, and $\hat{P}_{s,y} := \{(X_i, Y_i) \in \hat{P} : (X_i)_s = s, Y_i = y\}$. Then for $\delta > 0$, with respect to the random draw of $\hat{P}$ from $\mathcal{P}$, we have*

$$\Pr\left(\hat{L}_n - M\sqrt{\frac{\ln(2/\delta)}{2|\hat{P}_{s,y}|}} \le \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)] \le \hat{L}_n + M\sqrt{\frac{\ln(2/\delta)}{2|\hat{P}_{s,y}|}}\right) \ge 1 - \delta, \tag{38}$$

$$\Pr\left(\sqrt{s_n^2} - M\sqrt{\frac{2\ln(2/\delta)}{|\hat{P}_{s,y}| - 1}} \le \sqrt{\mathop{\mathbb{V}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]} \le \sqrt{s_n^2} + M\sqrt{\frac{2\ln(2/\delta)}{|\hat{P}_{s,y}| - 1}}\right) \ge 1 - \delta, \tag{39}$$

$$\Pr\left(\frac{|\hat{P}_{s,y}|}{|\hat{P}|} - \sqrt{\frac{\ln(2/\delta)}{2|\hat{P}|}} \le \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y] \le \frac{|\hat{P}_{s,y}|}{|\hat{P}|} + \sqrt{\frac{\ln(2/\delta)}{2|\hat{P}|}}\right) \ge 1 - \delta. \tag{40}$$

Proof of Lemma D.1. We can get Equation (39) according to Theorem 10 in [32]. Here, we will provide proofs for Equation (38) and Equation (40), respectively. The general idea is to use Hoeffding's inequality to get the high-confidence interval.

We will prove Equation (38) first. From Hoeffding's inequality, for all $t > 0$, we have:

$$\Pr\left(\left|\hat{L}_n - \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]\right| \ge t\right) \le 2\exp\left(-\frac{2|\hat{P}_{s,y}|^2 t^2}{|\hat{P}_{s,y}| M^2}\right) \tag{41}$$

Since we want to get an interval with confidence $1 - \delta$, we let $2\exp\left(-\frac{2|\hat{P}_{s,y}|^2 t^2}{|\hat{P}_{s,y}| M^2}\right) = \delta$, from which we can derive that

$$t = M\sqrt{\frac{\ln(2/\delta)}{2|\hat{P}_{s,y}|}} \tag{42}$$

Plugging Equation (42) into Equation (41), we can get:

$$\Pr\left(\hat{L}_n - M\sqrt{\frac{\ln(2/\delta)}{2|\hat{P}_{s,y}|}} \leq \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell\left(h_\theta(X), Y\right)] \leq \hat{L}_n + M\sqrt{\frac{\ln(2/\delta)}{2|\hat{P}_{s,y}|}}\right) \geq 1-\delta \tag{43}$$

Then we will prove Equation (40). From Hoeffding's inequality, for all $t > 0$, we have:

$$\Pr\left(\left|\frac{|\hat{P}_{s,y}|}{|\hat{P}|} - \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y]\right| \geq t\right) \leq 2\exp\left(-\frac{2|\hat{P}|^2 t^2}{|\hat{P}|}\right) \tag{44}$$

Since we want to get an interval with confidence $1-\delta$, we let $2\exp\left(-\frac{2|\hat{P}|^2 t^2}{|\hat{P}|}\right) = \delta$, from which we can derive that

$$t = \sqrt{\frac{\ln(2/\delta)}{2|\hat{P}|}} \tag{45}$$

Plugging Equation (45) into Equation (44), we can get:

$$\Pr\left(\frac{|\hat{P}_{s,y}|}{|\hat{P}|} - \sqrt{\frac{\ln(2/\delta)}{2|\hat{P}|}} \leq \Pr_{(X,Y)\sim\mathcal{P}}[X_s = s, Y = y] \leq \frac{|\hat{P}_{s,y}|}{|\hat{P}|} + \sqrt{\frac{\ln(2/\delta)}{2|\hat{P}|}}\right) \geq 1-\delta \tag{46}$$

□

D.2 Fairness Certification Statements with Finite Sampling

Theorem 5 (Thm. 2 with finite sampling)**.** *Given a distance bound $\rho > 0$ and any $\delta > 0$, the following constrained optimization, which is **convex**, when feasible, provides a fairness certificate for Problem 2 with probability at least $1 - 2SC\delta$:*

$$\max_{k_s, r_y, p_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \overline{E_{s,y}} \tag{47a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S} k_s = 1, \quad \sum_{y=1}^{C} r_y = 1, \quad k_s \geq 0 \quad \forall s \in [S], \quad r_y \geq 0 \quad \forall y \in [C], \tag{47b}$$

$$1 - \rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y} k_s r_y} \leq 0, \tag{47c}$$

$$\underline{p_{s,y}} \leq p_{s,y} \leq \overline{p_{s,y}}, \quad \forall s \in [S], \quad \forall y \in [C] \tag{47d}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C} p_{s,y} = 1 \tag{47e}$$

where $\overline{E_{s,y}} := \hat{L}_n + M\sqrt{\ln(2/\delta)/\left(2|\hat{P}_{s,y}|\right)}$, $\underline{p_{s,y}} := |\hat{P}_{s,y}|/|\hat{P}| - \sqrt{\ln(2/\delta)/\left(2|\hat{P}|\right)}$, $\overline{p_{s,y}} := |\hat{P}_{s,y}|/|\hat{P}| + \sqrt{\ln(2/\delta)/\left(2|\hat{P}|\right)}$ are constants computed with Lemma D.1.

Theorem 6. *If for any $s \in [S]$ and $y \in [Y]$, $H(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y}) \leq \bar{\gamma}_{s,y}$ and $0 \leq \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \ell(h_\theta(X), Y) \leq M$, given a distance bound $\rho > 0$ and any $\delta > 0$, for any region granularity $T \in \mathbb{N}_+$, the following expression provides a fairness certificate for Problem 1 with probability at least $1 - 3SC\delta$:*

$$\bar{\ell} = \max_{\{i_s\in[T]:s\in[S]\},\{j_y\in[T]:y\in[C]\}} \mathsf{C}\left(\left\{\left[\frac{i_s-1}{T}, \frac{i_s}{T}\right]\right\}_{s=1}^{S}, \left\{\left[\frac{j_y-1}{T}, \frac{j_y}{T}\right]\right\}_{y=1}^{C}\right), \text{ where} \tag{48}$$

$$\mathsf{C}\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right) = \max_{x_{s,y}, p_{s,y}} \sum_{s=1}^{S}\sum_{y=1}^{C}\left(\overline{k_s}\,\overline{r_y}\left(\overline{E_{s,y}} + \overline{C_{s,y}}\right)_+ + \underline{k_s}\,\underline{r_y}\left(\overline{E_{s,y}} + \overline{C_{s,y}}\right)_-\right.$$

$$+2\overline{k_s}\overline{r_y}\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{\overline{V_{s,y}}} - \underline{k_s}\underline{r_y}x_{s,y}(\underline{C_{s,y}})_+ - \overline{k_s}\overline{r_y}x_{s,y}(\underline{C_{s,y}})_-\Bigg) \tag{49a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S}\underline{k_s} \le 1, \quad \sum_{s=1}^{S}\overline{k_s} \ge 1, \quad \sum_{y=1}^{C}\underline{r_y} \le 1, \quad \sum_{y=1}^{C}\overline{r_y} \ge 1, \tag{49b}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}} \ge 1-\rho^2, \quad \left(1-\overline{\bar{\gamma}_{s,y}^2}\right)^2 \le x_{s,y} \le 1, \tag{49c}$$

$$\underline{p_{s,y}} \le p_{s,y} \le \overline{p_{s,y}}, \quad \sum_{s=1}^{S}\sum_{y=1}^{C}p_{s,y} = 1 \tag{49d}$$

where $(\cdot)_+ = \max\{\cdot,0\}$, $(\cdot)_- = \min\{\cdot,0\}$; $\underline{E_{s,y}} := \hat{L}_n - M\sqrt{\ln(2/\delta)/\left(2|\hat{P}_{s,y}|\right)}$, $\overline{E_{s,y}} := \hat{L}_n + M\sqrt{\ln(2/\delta)/\left(2|\hat{P}_{s,y}|\right)}$, $\underline{V_{s,y}} = \left(\sqrt{s_n^2} - M\sqrt{2\ln(2/\delta)/\left(|\hat{P}_{s,y}|-1\right)}\right)^2$, $\overline{V_{s,y}} = \left(\sqrt{s_n^2} + M\sqrt{2\ln(2/\delta)/\left(|\hat{P}_{s,y}|-1\right)}\right)^2$, $\underline{p_{s,y}} := |\hat{P}_{s,y}|/|\hat{P}| - \sqrt{\ln(2/\delta)/\left(2|\hat{P}|\right)}$, $\overline{p_{s,y}} := |\hat{P}_{s,y}|/|\hat{P}| + \sqrt{\ln(2/\delta)/\left(2|\hat{P}|\right)}$ *computed with Lemma D.1, and* $\underline{C_{s,y}} = M - \overline{E_{s,y}} - \overline{V_{s,y}}/(M-\overline{E_{s,y}})$, $\overline{C_{s,y}} = M - \underline{E_{s,y}} - \underline{V_{s,y}}/(M-\underline{E_{s,y}})$, $\overline{\bar{\gamma}_{s,y}^2} = 1-(1+(M-\underline{E_{s,y}})^2/\overline{V_{s,y}})^{-\frac{1}{2}}$. *Equation* (48) *only takes* **C***'s value when it is feasible, and each* **C** *queried by Equation* (48) *is a* ***convex optimization***.

D.3 Proofs of Fairness Certification with Finite Sampling

High-Level Illustration. We use Hoeffding's inequality to bound the finite sampling error of statistics and add the high confidence box constraints to the optimization problems, which can still be proved to be convex.

Proof of Thm. 5. The proof of Thm. 5 is composed of two parts: (1) the optimization problem provides a fairness certificate for Problem 2; (2) the optimization problem is convex.

(1) We prove that Thm. 5 provides a fairness certificate for Problem 2 in this part. Since Thm. 2 provides a fairness certificate for Problem 2, we only need to prove: (a) the feasible region of the optimization problem in Thm. 2 is a subset of the feasible region of the optimization problem in Thm. 5, and (b) the optimization objective in Thm. 2 can be upper bounded by that in Thm. 5.

To prove (a), we first equivalently transform the optimization problem in Thm. 2 into the following optimization problem by adding p_{sy} to the decision variables:

$$\max_{k_s,r_y,p_{s,y}} \quad \sum_{s=1}^{S}\sum_{y=1}^{C}k_s r_y E_{s,y} \tag{50a}$$

$$\text{s.t.} \quad \sum_{s=1}^{S}k_s = 1, \quad \sum_{y=1}^{C}r_y = 1, \quad k_s \ge 0 \quad \forall s\in[S], \quad r_y \ge 0 \quad \forall y \in [C], \tag{50b}$$

$$1-\rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}k_s r_y} \le 0, \tag{50c}$$

$$p_{s,y} = |\hat{P}_{s,y}|/|\hat{P}|, \quad \forall s\in[S], \quad \forall y\in[C] \tag{50d}$$

$$\sum_{s=1}^{S}\sum_{y=1}^{C}p_{s,y} = 1 \tag{50e}$$

For decision variables $k_{s,y}$ and $r_{s,y}$, optimization 47 and 50 has the same constraints (Equation (47b) and Equation (50b)). For decision variables $p_{s,y}$, the feasible region of $p_{s,y}$ in

optimization 47 (decided by Equations (47d) and (47e)) is a subset of the feasible region of $p_{s,y}$ in optimization 50 (decided by Equations (50d) and (50e)), since $\underline{p_{s,y}} \leq |\hat{P}_{s,y}|/|\hat{P}| \leq \overline{p_{s,y}}$. Therefore, the feasible region with respect to $k_{s,y}$, $r_{s,y}$, and $p_{s,y}$ of the optimization problem in Thm. 2 is a subset of that in Thm. 5.

To prove (b), we only need to show that the objective in Equation (47a) can be upper bounded by the objective in Equation (50a). The statement $\sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y E_{s,y} \leq \sum_{s=1}^{S}\sum_{y=1}^{C} k_s r_y \overline{E_{s,y}}$ consistently holds because $E_{s,y} \leq \overline{E_{s,y}}$ and $k_s, r_y \geq 0$.

Combining the proofs of (a) and (b), we prove that Thm. 5 provides a fairness certificate for Problem 2.

(2) Inspecting that the objective and all the constraints in optimization problem in Equation (47) are linear with respect to k_s, r_y, $p_{s,y}$ but the one in Equation (47c). Therefore, we only need to prove that the following constraint is convex with respect to k_s, r_y, $p_{s,y}$:

$$1 - \rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y} \leq 0 \tag{51}$$

We define a function f with respect to vector $\mathbf{p} := [p_{s,y}]_{s\in[S],y\in[C]}$: $f(\mathbf{p}) = 1 - \rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y} k_s r_y}$. Then we can derive that:

$$\frac{\partial^2 f}{\partial \mathbf{p}^2} = \sum_{s=1}^{S}\sum_{y=1}^{C} \frac{\sqrt{k_s r_y}}{4} {p_{s,y}}^{-\frac{3}{2}} \geq 0 \tag{52}$$

Therefore, the function f is convex with respect to $p_{s,y}$. Similarly, we can prove the convexity with respect to k_s and r_y. Finally, we can conclude that the constraint in Equation (51) is convex with respect to k_s, r_y, $p_{s,y}$ and the optimization problem defined in Thm. 5 is convex.

Since we use the union bound to bound $E_{s,y}$ and $p_{s,y}$ for all $s \in [S], y \in [C]$ simultaneously, the confidence is $1 - 2SC\delta$. □

Proof of Thm. 6. The proof of Thm. 6 includes two parts: (1) the optimization problem provides a fairness certificate for Problem 1; (2) each **C** queried by Equation (48) is a convex optimization.

(1) Since Thm. 3 provides a fairness certificate for Problem 1, we only need to prove: (a) the feasible region of the optimization problem in Thm. 3 is a subset of that in Thm. 6, and (b) the optimization objective in Thm. 3 can be upper bounded by that in Thm. 6.

To prove (a), we first equivalently transform the optimization problem in Thm. 3 into the following optimization problem by adding p_{sy} to the decision variables:

$$\begin{aligned}
&\mathbf{C}\left(\{[\underline{k_s}, \overline{k_s}]\}_{s=1}^{S}, \{[\underline{r_y}, \overline{r_y}]\}_{y=1}^{C}\right) = \max_{x_{s,y}, p_{s,y}} \sum_{s=1}^{S}\sum_{y=1}^{C} \Big(\overline{k_s}\,\overline{r_y}\,(E_{s,y} + C_{s,y})_+ + \underline{k_s}\,\underline{r_y}\,(E_{s,y} + C_{s,y})_- \\
&+2\overline{k_s}\,\overline{r_y}\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}} - \underline{k_s}\,\underline{r_y} x_{s,y}(C_{s,y})_+ - \overline{k_s}\,\overline{r_y} x_{s,y}(C_{s,y})_-\Big) && \text{(53a)}\\
&\text{s.t.} \quad \sum_{s=1}^{S} \underline{k_s} \leq 1, \quad \sum_{s=1}^{S} \overline{k_s} \geq 1, \sum_{y=1}^{C} \underline{r_y} \leq 1, \quad \sum_{y=1}^{C} \overline{r_y} \geq 1, && \text{(53b)}\\
&\sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{p_{s,y}\overline{k_s}\,\overline{r_y} x_{s,y}} \geq 1 - \rho^2, \quad \left(1 - \bar{\gamma}_{s,y}^2\right)^2 \leq x_{s,y} \leq 1, && \text{(53c)}\\
&p_{s,y} = |\hat{P}_{s,y}|/|\hat{P}|, \quad \forall s \in [S], \quad \forall y \in [C] && \text{(53d)}\\
&\sum_{s=1}^{S}\sum_{y=1}^{C} p_{s,y} = 1 && \text{(53e)}
\end{aligned}$$

For decision varibales $x_{s,y}$, since $\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}} \leq \sqrt{\overline{p_{s,y}}\overline{k_s}\overline{r_y}x_{s,y}}$ and $\left(1-\bar{\gamma}_{s,y}^2\right)^2 \geq \left(1-\overline{\bar{\gamma}_{s,y}^2}\right)^2$, the feasible region of $x_{s,y}$ in Equation (53) is a subset of that in Equation (49). For decision variables $p_{s,y}$, since $\underline{p_{s,y}} \leq |\hat{P}_{s,y}|/|\hat{P}| \leq \overline{p_{s,y}}$, the feasible region of $p_{s,y}$ in Equation (53) is also a subset of that in Equation (49). Therefore, the feasible region of the optimization problem in Thm. 3 is a subset of that in Thm. 6.

To prove (b), we only need to show that the objective in Equation (53a) can be upper bounded by the objective in Equation (49a). Since $\underline{k_s}, \overline{k_s}, \underline{r_y}, \overline{r_y} \geq 0$ and $0 \leq x_{s,y} \leq 1$ hold, we can observe that $\forall s \in [S], \forall y \in [C]$,

$$\overline{k_s}\overline{r_y}\left(E_{s,y}+C_{s,y}\right)_+ + \underline{k_s}\underline{r_y}\left(E_{s,y}+C_{s,y}\right)_- + 2\overline{k_s}\overline{r_y}\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{V_{s,y}} - \underline{k_s}\underline{r_y}x_{s,y}(C_{s,y})_+ - \overline{k_s}\overline{r_y}x_{s,y}(C_{s,y})_- \leq \overline{k_s}\overline{r_y}\left(\overline{E_{s,y}}+\overline{C_{s,y}}\right)_+ + \underline{k_s}\underline{r_y}\left(\overline{E_{s,y}}+\overline{C_{s,y}}\right)_- + 2\overline{k_s}\overline{r_y}\sqrt{x_{s,y}(1-x_{s,y})}\sqrt{\overline{V_{s,y}}} - \underline{k_s}\underline{r_y}x_{s,y}(\underline{C_{s,y}})_+ - \overline{k_s}\overline{r_y}x_{s,y}(\underline{C_{s,y}})_-$$

Therefore, we prove that the optimization in Thm. 6 provides a fairness certificate for Problem 1.

(2) We will prove that each **C** queried by Equation (48) is a convex optimization with respect to decision variables $x_{s,y}$ and $p_{s,y}$ in this part. In the proof of Thm. 3, we provide the proof of convexity with respect to $x_{s,y}$, so we only need to prove that the optimization problem is convex with respect to $p_{s,y}$. We can observe that the constraints of $p_{s,y}$ in Equation (49d) is linear, and thus we only need to prove that $\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}} \geq 1-\rho^2$ (the constraint in Equation (49c)) is convex with respect to $p_{s,y}$. Here, we define a function f with respect to vector $\boldsymbol{p} := [p_{s,y}]_{s\in[S],y\in[C]}$: $f(\boldsymbol{p}) = 1-\rho^2-\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}\overline{k_s}\overline{r_y}}$. Then we can derive that:

$$\left(\frac{\partial^2 f}{\partial \boldsymbol{p}^2}\right)_{sy,s'y'} = \sum_{s=1}^{S}\sum_{y=1}^{C}\frac{\sqrt{\overline{k_s}\overline{r_y}}}{4}{p_{s,y}}^{-\frac{3}{2}}\cdot\mathbb{I}[s=s',y=y'] \geq 0 \tag{55}$$

Thus, the function f is convex and $f(\boldsymbol{p}) \leq 0$ defines a convex set with respect to $p_{s,y}$. Then, we prove that the constraint $\sum_{s=1}^{S}\sum_{y=1}^{C}\sqrt{p_{s,y}\overline{k_s}\overline{r_y}x_{s,y}} \geq 1-\rho^2$ is convex with respect to $p_{s,y}$.

Since we use the union bound to bound $E_{s,y}$, $V_{s,y}$ and $p_{s,y}$ for all $s \in [S], y \in [C]$ simultaneously, the confidence is $1-3SC\delta$. □

E Experiments

E.1 Datasets

We validate our certified fairness on *six* real-world datasets: Adult [3], Compas [2], Health [19], Lawschool [48], Crime [3], and German [3]. All the used datasets contain categorical data.

In Adult dataset, we have 14 attributes of a person as input and try to predict whether the income of the person is over 50k \$/year. The sensitive attribute in Adult is selected as the sex.

In Compas dataset, given the attributes of a criminal defendent, the task is to predict whether he/she will re-offend in two years. The sensitive attribute in Compas is selected as the race.

In Health dataset, given the physician records and and insurance claims of the patients, we try to predict ten-year mortality by binarizing the Charlson Index, taking the median value as a cutoff. The sensitive attribute in Health is selected as the age.

In Lawschool dataset, we try to predict whether a student passes the exam according to the appication records of different law schools. The sensitive attribute in Lawschool is the race.

In Crime dataset, we try to predict whether a specific community is above or below the median number of violent crimes per population. The sensitive attribute in Crime is selected as the race.

In German dataset, each person is classified as good or bad credit risks according to the set of attributes. The sensitive attribute in German is selected as the sex.

Following [39], we consider the scenario where sensitive attributes and labels take binary values, and we also follow their standard data processing steps: (1) normalize the numerical values of all attributes with the mean value 0 and variance 1, (2) use one-hot encodings to represent categorical attributes, (3) drop instances and attributes with missing values, and (4) split the datasets into training set, validation set, and test set.

Training Details. We directly train a ReLU network composed of two hidden layers on the training set of six datasets respectively following the setting in [39]. Concretely, we train the models for 100 epochs with a batch size of 256. We adopt the binary cross-entropy loss and use the Adam optimizer with weight decay 0.01 and dynamic learning rate scheduling (ReduceLROnPlateau in [35]) based on the loss on the validation set starting at 0.01 with the patience of 5 epochs.

E.2 Fair Base Rate Distribution Generation Protocol

To evaluate how well our certificates capture the fairness risk in practice, we compare our certification bound with the empirical loss evaluated on randomly generated $30,000$ fairness constrained distributions $\mathcal{Q}$ shifted from $\mathcal{P}$. Now, we introduce the protocols to generate fairness distributions $\mathcal{Q}$ for sensitive shifting and general shifting, respectively. Note that the protocols are only valid when the sensitive attributes and labels take binary values.

Fair base rate distributions generation steps in the *sensitive shifting* scenario:

(1) Sample the proportions of subpopulations of the generated distribution $q_{0,0}, q_{0,1}, q_{1,0}, q_{1,1}$: uniformly sample two real values in the interval $[0, 1]$, and do the assignment: $q_{0,0} := kr$, $q_{0,1} := k(1-r)$, $q_{1,0} := (1-k)r$, $q_{1,1} := (1-k)(1-r)$.

(2) Determine the sample size of every subpopulation: first determine the subpopulation which requires the largest sample size, use all the samples in that subpopulation, and then calculate the sample size in other subpopulations according to the proportions.

(3) Uniformly sample in each subpopulation based on the sample size.

(4) Calculate the Hellinger distance $H(\mathcal{P}, \mathcal{Q}) = \sqrt{1 - \sum_{s=0}^{1} \sum_{y=0}^{1} \sqrt{p_{s,y}} \sqrt{q_{s,y}}}$. Suppose that the support of $\mathcal{P}$ and $\mathcal{Q}$ is $\mathcal{X} \times \mathcal{Y}$ and the densities of $\mathcal{P}$ and $\mathcal{Q}$ with respect to a suitable measure are $f_{\mathcal{P}}$ and $f_{\mathcal{Q}}$, respectively. Since we consider sensitive shifting here, we have $f_{\mathcal{Q}_{s,y}} = \lambda_{s,y} f_{\mathcal{P}_{s,y}}, \quad s, y \in \{0, 1\}$ where $\lambda_{s,y}$ is a scalor. The derivation of the distance calculation formula is shown as follows,

$$H^2(\mathcal{P}, \mathcal{Q}) = 1 - \iint_{\mathcal{X} \times \mathcal{Y}} \sqrt{f_{\mathcal{P}}(x, y)} \sqrt{f_{\mathcal{Q}}(x, y)} \mathrm{d}x \mathrm{d}y \tag{56a}$$

$$= 1 - \sum_{s=0}^{1} \sum_{y=0}^{1} \iint_{f_{\mathcal{P}_{s,y}}(x,y) > 0} \sqrt{f_{\mathcal{P}_{s,y}}(x, y)} \sqrt{\lambda_{s,y} f_{\mathcal{P}_{s,y}}} \mathrm{d}x \mathrm{d}y \tag{56b}$$

$$= 1 - \sum_{s=0}^{1} \sum_{y=0}^{1} \sqrt{\lambda_{s,y}} \iint_{f_{\mathcal{P}_{s,y}}(x,y) > 0} f_{\mathcal{P}_{s,y}}(x, y) \mathrm{d}x \mathrm{d}y \tag{56c}$$

$$= 1 - \sum_{s=0}^{1} \sum_{y=0}^{1} \sqrt{\lambda_{s,y}} p_{s,y} \tag{56d}$$

$$= 1 - \sum_{s=0}^{1} \sum_{y=0}^{1} \sqrt{p_{s,y}} \sqrt{q_{s,y}}. \tag{56e}$$

Fair base rate distribution generation steps in the *general shifting* scenario:

(1) Construct a data distribution $\mathcal{Q}'$ that is disjoint with the training data distribution $\mathcal{P}$ by changing the distribution of non-sensitive values given the sensitive attributes and labels.

(2) Sample mixing parameters $\alpha_{s,y}$ and $\alpha'_{s,y}$ in the interval $[0,1]$ satisfying $\frac{p_{00}\alpha_{00}+q_{00}\alpha'_{00}}{p_{01}\alpha_{01}+q_{01}\alpha'_{01}} = \frac{p_{10}\alpha_{10}+q_{10}\alpha'_{10}}{p_{11}\alpha_{11}+q_{11}\alpha'_{11}}$ (base rate parity) and $p_{00}\alpha_{00} + q_{00}\alpha'_{00} + p_{01}\alpha_{01} + q_{01}\alpha'_{01} + p_{10}\alpha_{10} + q_{10}\alpha'_{10} + p_{11}\alpha_{11} + q_{11}\alpha'_{11} = 1$.

(3) Determine the proportion of every subpopulation in distribution $\mathcal{Q}$: $q_{s,y} := \alpha_{s,y}p_{s,y} + \alpha'_{s,y}q'_{s,y}, \quad s,y \in \{0,1\}$.

(4) Determine the sample size of every subpopulation in $\mathcal{P}$ and $\mathcal{Q}'$: first determine the subpopulation which requires the largest sample size, use all the samples in that subpopulation, and then calculate the sample size in other subpopulations according to the proportions.

(5) Calculate the Hellinger distance between distribution $\mathcal{P}$ and $\mathcal{Q}$: $H(\mathcal{P},\mathcal{Q}) = \sqrt{1-\sum_{s=0}^{1}\sum_{y=0}^{1}\sqrt{\alpha_{s,y}}p_{s,y}}$. Suppose that the support of $\mathcal{P}$ and $\mathcal{Q}$ is $\mathcal{X}\times\mathcal{Y}$ and the densities of $\mathcal{P}$ and $\mathcal{Q}$ with respect to a suitable measure are $f_{\mathcal{P}}$ and $f_{\mathcal{Q}}$, respectively. The derivation of the distance calculation formula is shown as follows,

$$H^2(\mathcal{P},\mathcal{Q}) = 1 - \iint_{\mathcal{X}\times\mathcal{Y}} \sqrt{f_{\mathcal{P}}(x,y)}\sqrt{f_{\mathcal{Q}}(x,y)}\mathrm{d}x\mathrm{d}y \tag{57a}$$

$$= 1 - \iint_{\mathcal{X}\times\mathcal{Y}} \sqrt{\sum_{s=0}^{1}\sum_{y=0}^{1} f_{\mathcal{P}_{s,y}}(x,y)}\sqrt{\sum_{s=0}^{1}\sum_{y=0}^{1}\Big(\alpha_{s,y}f_{\mathcal{P}_{s,y}}(x,y) + \alpha'_{s,y}f_{\mathcal{Q}'_{s,y}}(x,y)\Big)}\mathrm{d}x\mathrm{d}y \tag{57b}$$

$$= 1 - \sum_{s=0}^{1}\sum_{y=0}^{1}\sqrt{\alpha_{s,y}}\iint_{f_{\mathcal{P}_{s,y}}(x,y)>0} f_{\mathcal{P}_{s,y}}(x,y)\sqrt{1+\frac{\alpha'_{s,y}f_{\mathcal{Q}_{s,y}}(x,y)}{\alpha_{s,y}f_{\mathcal{P}_{s,y}}(x,y)}}\mathrm{d}x\mathrm{d}y \tag{57c}$$

$$= 1 - \sum_{s=0}^{1}\sum_{y=0}^{1}\sqrt{\alpha_{s,y}}\iint_{f_{\mathcal{P}_{s,y}}(x,y)>0} f_{\mathcal{P}_{s,y}}(x,y)\mathrm{d}x\mathrm{d}y \tag{57d}$$

$$= 1 - \sum_{s=0}^{1}\sum_{y=0}^{1}\sqrt{\alpha_{s,y}}p_{s,y}. \tag{57e}$$

E.3 Implementation Details of Our Fairness Certification

We conduct vanilla training and then calculate our certified fairness according to our certification framework. Concretely, in the training step, we train a ReLU network composed of 2 hidden layers of size 20 for 100 epochs with binary cross entropy loss (BCE loss) using an Adam optimizer. The initial learning rate is 0.05 for Crime and German datasets, while for other datasets, the initial learning rate is set 0.001. We reduce the learning rate with a factor of 0.5 on the plateau measured by the loss on the validation set with patience of 5 epochs. In the fairness certification step, we set the region granularity to be 0.005 for certification in the general shifting scenario. We use 90% confidence interval when considering finite sampling error. The codes we used follow the MIT license. All experiments are conducted on a 1080 Ti GPU with 11,178 MB memory.

E.4 Implementation Details of WRM

The optimization problem of tackling distributional robustness is formulated as:

$$\max_{\mathcal{Q}} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X),Y)] \quad \text{s.t.} \quad \mathrm{dist}(\mathcal{P},\mathcal{Q}) \le \rho \tag{58}$$

where $\mathrm{dist}(\cdot,\cdot)$ is a predetermined distribution distance metric. Note that the optimization is the same as our certified fairness optimization in Problem 1 except for the fairness constraint.

WRM [43] proposes to use the dual reformulation of the Wasserstein worst-case risk to provide the distributional robustness certificate, which is formulated in the following proposition.

Proposition 2 ([43], Proposition 1)**.** *Let $\ell : \Theta\times\mathcal{Z}\to\mathbb{R}$ and $c : \Theta\times\mathcal{Z}\to\mathbb{R}_+$ be continuous and let $\phi_\gamma(\theta;z_0) := \sup_{z\in\mathcal{Z}}\{\ell(z;\theta)-\gamma c(z;\theta)\}$. Then, for any distribution $\mathcal{P}$ and any $\rho>0$,*

$$\sup_{\mathcal{Q}:W_c(\mathcal{P},\mathcal{Q})\le\rho} \mathbb{E}_{\mathcal{Q}}[\ell(\theta;Z)] = \inf_{\gamma\ge 0}\{\gamma\rho + \mathbb{E}_{\mathcal{P}}[\phi_\gamma(\theta;Z)]\} \tag{59}$$

where $W_c(\mathcal{P}, \mathcal{Q}) := \inf_{\pi \in \Pi(\mathcal{P},\mathcal{Q})} \int_{\mathcal{Z}} c(z, z') d\pi(z, z')$ *is the 1-Wasserstein distance between* $\mathcal{P}$ *and* $\mathcal{Q}$.

One requirement for the certificate to be tractable is that the surrogate function ϕ_γ is concave with respect to Z, which holds when γ is larger than the Lipschitz constant L of the gradient of ℓ with respect to Z. Since we use the ELU network with JSD loss, we can efficiently calculate γ iteratively as shown in Appendix D of [47].

We select Gaussian mixture data for fair comparison. The Gaussian mixture data can be formulated as $P(x|\theta) = \sum_{k=1}^{K} \alpha_k \phi(x|\theta_k)$ where K is the number of Gaussian data, α_k is the proportion of the k-th Gaussian, and $\theta_k = (\mu_k, \sigma_k^2)$. In our evaluation, we use 2-dimension Gaussian and mixture data composed of 2 Gaussian ($K = 2$) labeled 0 and 1, respectively. Concretely, we let $\mu_1 = (-2, -0.5), \sigma_1 = 1.0, \alpha_1 = 0.5$ and $\mu_2 = (2, 0.5), \sigma_2 = 1.0, \alpha_2 = 0.5$. The second dimension of input vector is selected as the sensitive attribute X_s, and the base rate constraint becomes: $\Pr(Y = 0|X_s < 0) = \Pr(Y = 1|X_s > 0)$. Given the perturbation $\delta \in \mathbb{R}^2$ that induces $X \mapsto X + \delta$, the Wasserstein distance and Hellinger distance can be formulated as follows:

$$W_2(\mathcal{P}, \mathcal{Q}) = \|\delta\|_2, \quad H(\mathcal{P}, \mathcal{Q}) = \sqrt{1 - e^{-\|\delta\|_2^2/8}}. \tag{60}$$

E.5 More Results of Certified Fairness with Sensitive Shifting and General shifting

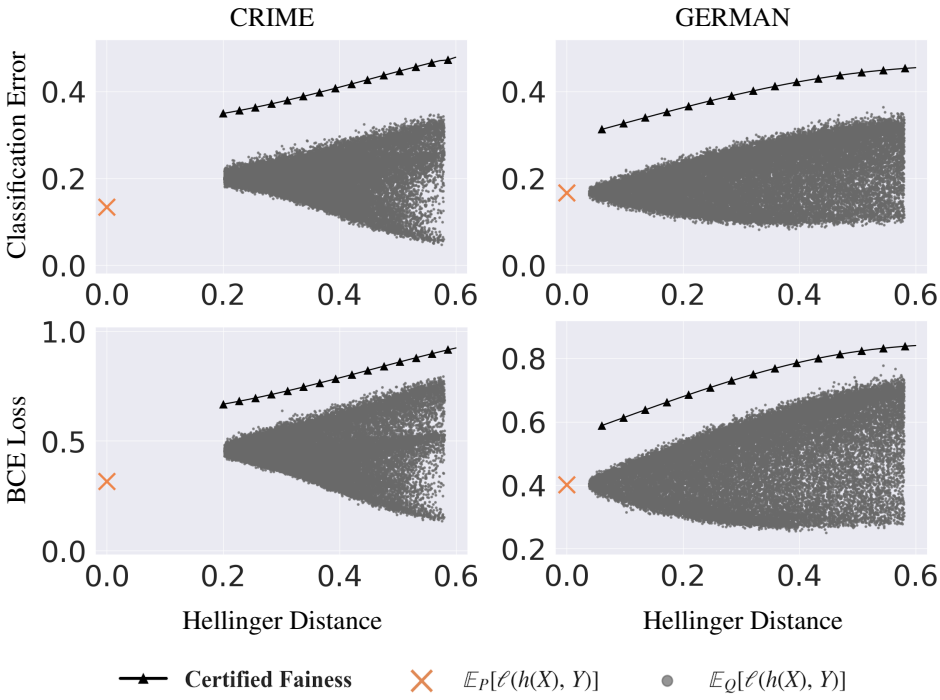


Figure 4: Certified fairness with sensitive shifting on Crime and German.

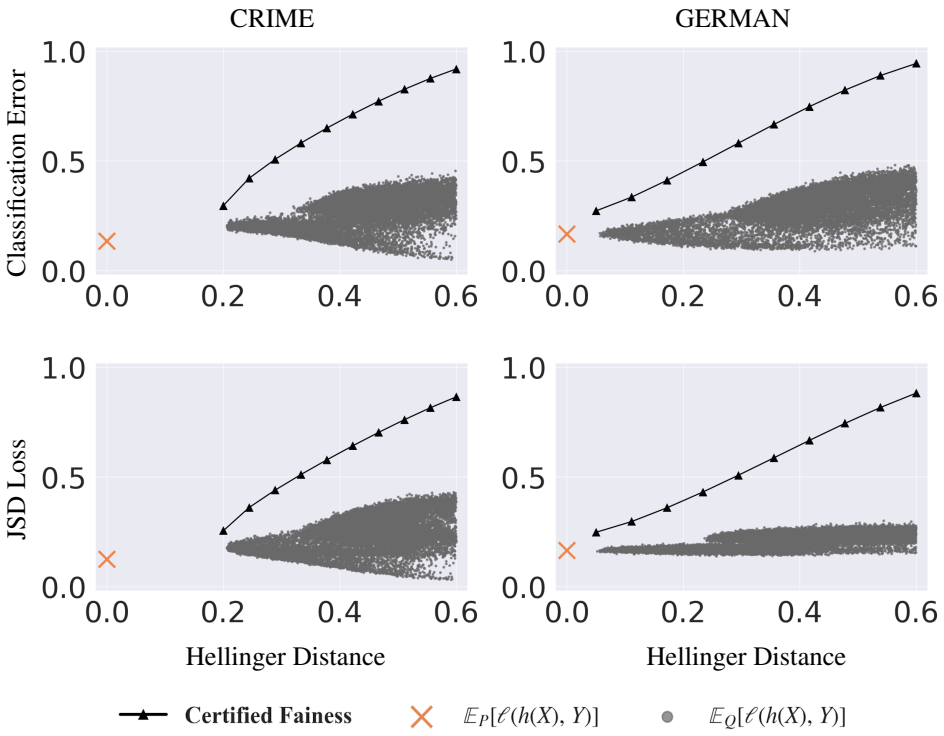


Figure 5: Certified fairness with general shifting on Crime and German.

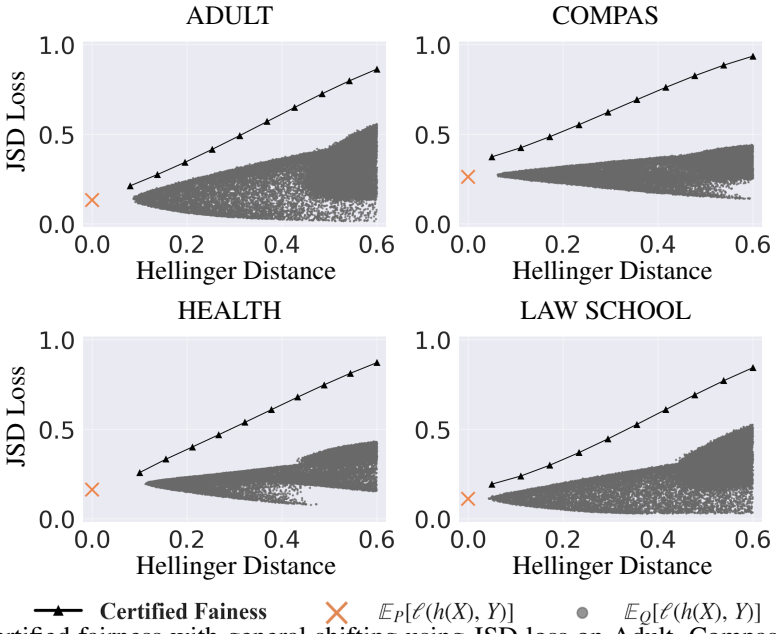


Figure 6: Certified fairness with general shifting using JSD loss on Adult, Compas, Health, and Lawschool.

E.6 More Results of Certified Fairness with Additional Non-Skewness Constraints

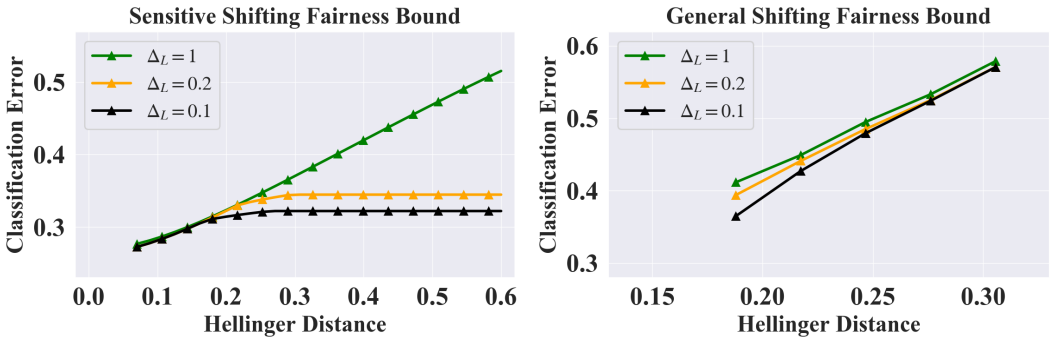


Figure 7: Certified fairness upper bounds with additional non-skewness constraints of labels on Adult. ($|\Pr_{(X,Y)\sim P}[Y=0] - \Pr_{(X,Y)\sim P}[Y=1]| \leq \Delta_L$)

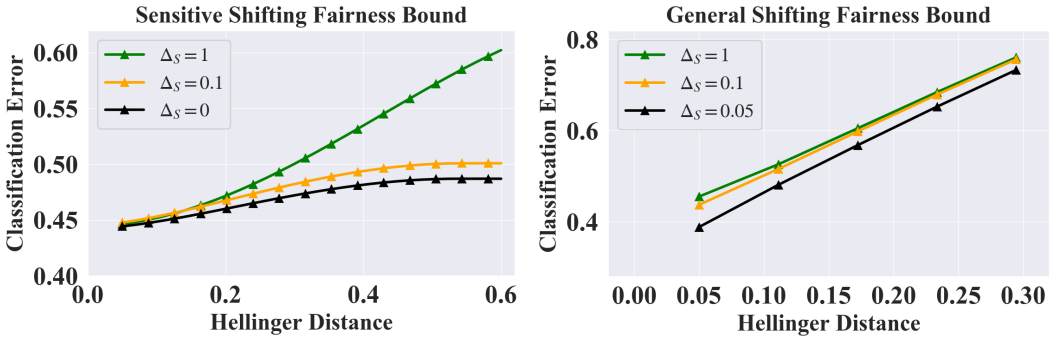


Figure 8: Certified fairness upper bounds with additional non-skewness constraints of sensitive attributes on Compas. ($|\Pr_{(X,Y)\sim P}[X_s=0] - \Pr_{(X,Y)\sim P}[X_s=1]| \leq \Delta_s$)

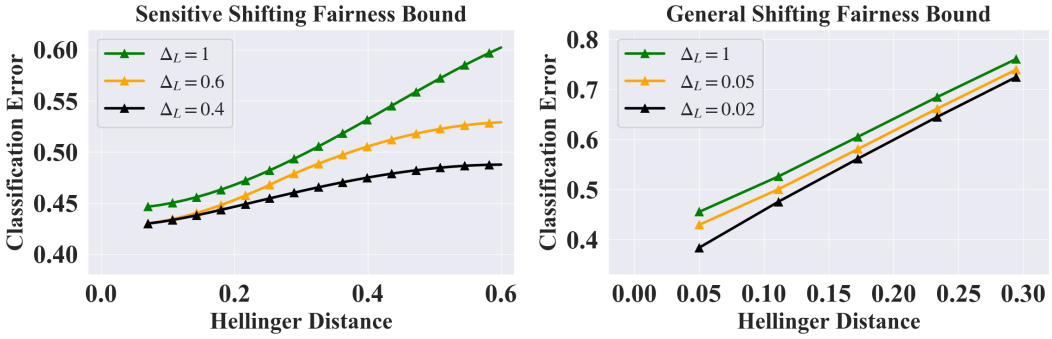


Figure 9: Certified fairness upper bounds with additional non-skewness constraints of labels on Compas. ($|\Pr_{(X,Y)\sim P}[Y=0] - \Pr_{(X,Y)\sim P}[Y=1]| \leq \Delta_L$)

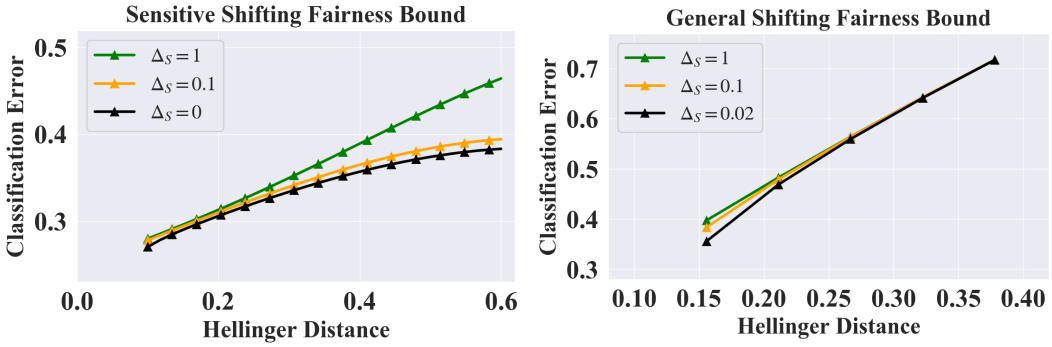


Figure 10: Certified fairness upper bounds with additional non-skewness constraints of sensitive attributes on Health. ($|\Pr_{(X,Y)\sim P}[X_s = 0] - \Pr_{(X,Y)\sim P}[X_s = 1]| \leq \Delta_s$)

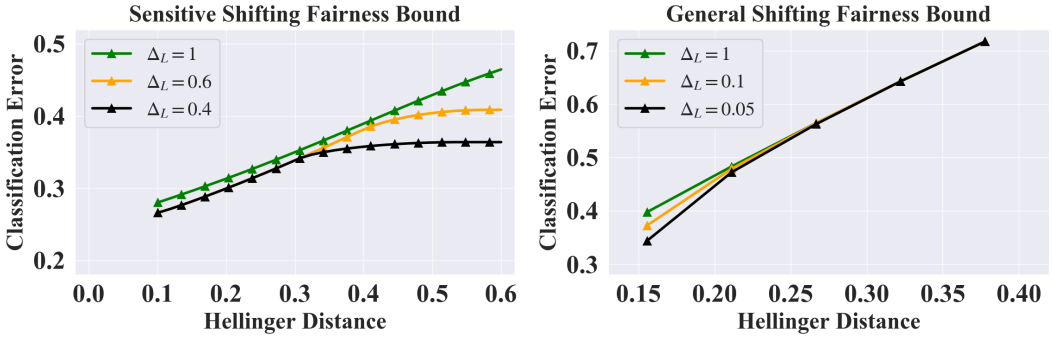


Figure 11: Certified fairness upper bounds with additional non-skewness constraints of labels on Health. ($|\Pr_{(X,Y)\sim P}[Y = 0] - \Pr_{(X,Y)\sim P}[Y = 1]| \leq \Delta_L$)

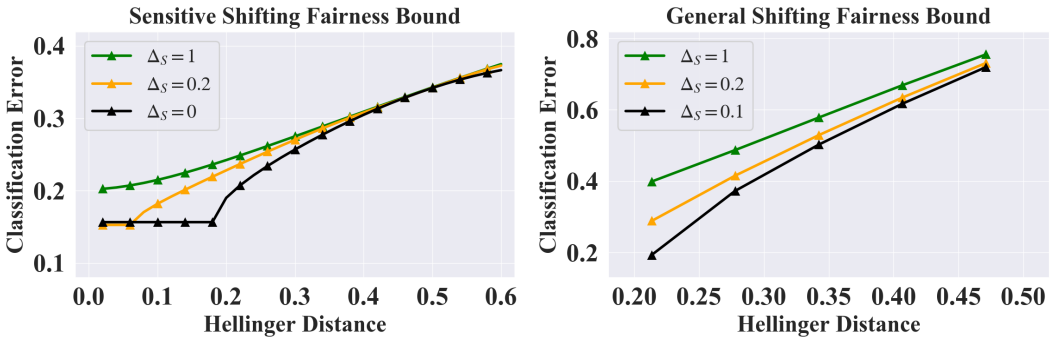


Figure 12: Certified fairness upper bounds with additional non-skewness constraints of sensitive attributes on Lawschool. ($|\Pr_{(X,Y)\sim P}[X_s = 0] - \Pr_{(X,Y)\sim P}[X_s = 1]| \leq \Delta_s$)

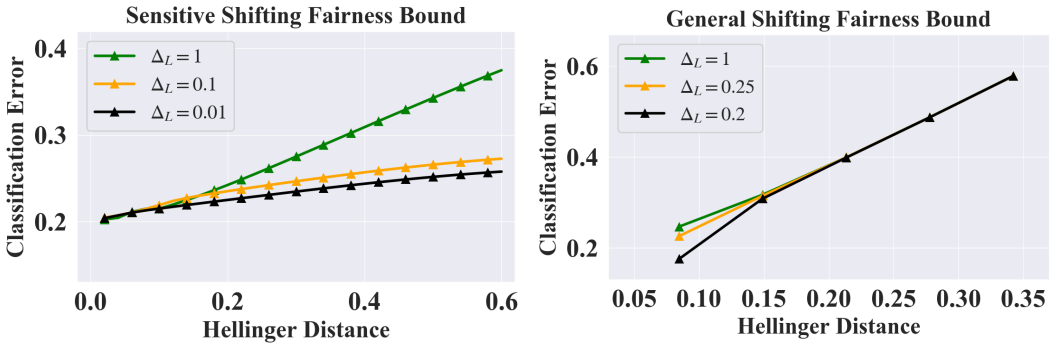


Figure 13: Certified fairness upper bounds with additional non-skewness constraints of labels on Lawschool. ($|\Pr_{(X,Y)\sim P}[Y=0] - \Pr_{(X,Y)\sim P}[Y=1]| \leq \Delta_L$)

E.7 Fair Classifier Achieves High Certified Fairness

We compare the fairness certificate of the vanilla model and the perfectly fair model on Adult dataset to demonstrate that our defined certified fairness in Problem 1 and Problem 2 can indicate the fairness in realistic scenarios. In Adult dataset, we have 14 attributes of a person as input and try to predict whether the income of the person is over 50k \$/year. The sensitive attribute in Adult is selected as the sex. We consider four subpopulations in the scenario: 1) male with salary below 50k, 2) male with salary above 50k, 3) female with salary below 50k, and 4) female with salary above 50k. We take the overall 0-1 error as the loss. The vanilla model is real, and trained with standard training loss on the Adault dataset. The perfectly fair model is hypothetical and simulated by enforcing the loss within each subpopulation to be the same as the vanilla trained classifier's overall expected loss for fair comparison with the vanilla model. From the experiment results in Table 1 and Table 2, we observe that our fairness certificates correlate with the actual fairness level of the model and verify that our certificates can be used as model's fairness indicator: the certified fairness of perfectly fair models are consistently higher than those for the unfair model, for both the general shifting scenario and the sensitive shifting scenario. These findings demonstrate the practicality of our fairness certification.

Table 1: Comparison of the fairness certificate of the vanilla model (an "unfair" model) and the perfectly fair model (a "fair" model) for sensitive shifting. 0-1 error is selected as the loss in the evaluation.

Hellinger Distance ρ	0.1	0.2	0.3	0.4	0.5
Vanilla Model Fairness Certificate	0.182	0.243	0.297	0.349	0.397
Fair Model Fairness Certificate	0.148	0.148	0.148	0.148	0.148

Table 2: Comparison of the fairness certificate of the vanilla model (an "unfair" model) and the perfectly fair model (a "fair" model) for general shifting. 0-1 error is selected as the loss in the evaluation.

Hellinger Distance ρ	0.1	0.2	0.3	0.4	0.5
Vanilla Model Fairness Certificate	0.274	0.414	0.559	0.701	0.828
Fair Model Fairness Certificate	0.266	0.407	0.553	0.695	0.824