Certifying Some Distributional Fairness with Subpopulation Decomposition

Mintong Kang*[1], Linyi Li*[1], Maurice Weber[2], Yang Liu[3], Ce Zhang[2], Bo Li[1]

[1]University of Illinois Urbana-Champaign [2]ETH Zürich [3]University of California, Santa Cruz

arxiv.org/abs/2205.15494

Introduction

- ML systems may be biased towards particular groups
- Existing approaches mainly **evaluate** fairness
- ➢ Important & challenging to rigorously **certify** fairness, which is our focus

Main Contributions:

- **We formulate** certified fairness problem of an end-to-end ML model
- **We propose an effective fairness certification framework** that **for the first time** solves this certified fairness problem by subpopulation decomposition
- **We evaluate** our framework on **6** real-world datasets to show its tightness and scalability

Core Methodology: Subpopulation Decomposition

Decompose according to sensitve attribute X_s and label Y

$$\mathcal{P} = \sum_{s=1}^{S}\sum_{y=1}^{C} \Pr_{\mathcal{P}}[X_s = s, Y = y] \cdot \mathcal{P}_{s,y},$$

$$\mathcal{Q} = \sum_{s=1}^{S}\sum_{y=1}^{C} \Pr_{\mathcal{Q}}[X_s = s, Y = y] \cdot \mathcal{Q}_{s,y}$$

Problem Formulation

Given model $h_\theta(\cdot)$, compute an upper bound of its expected loss on a fair test distribution $\mathcal{Q}$, i.e.,

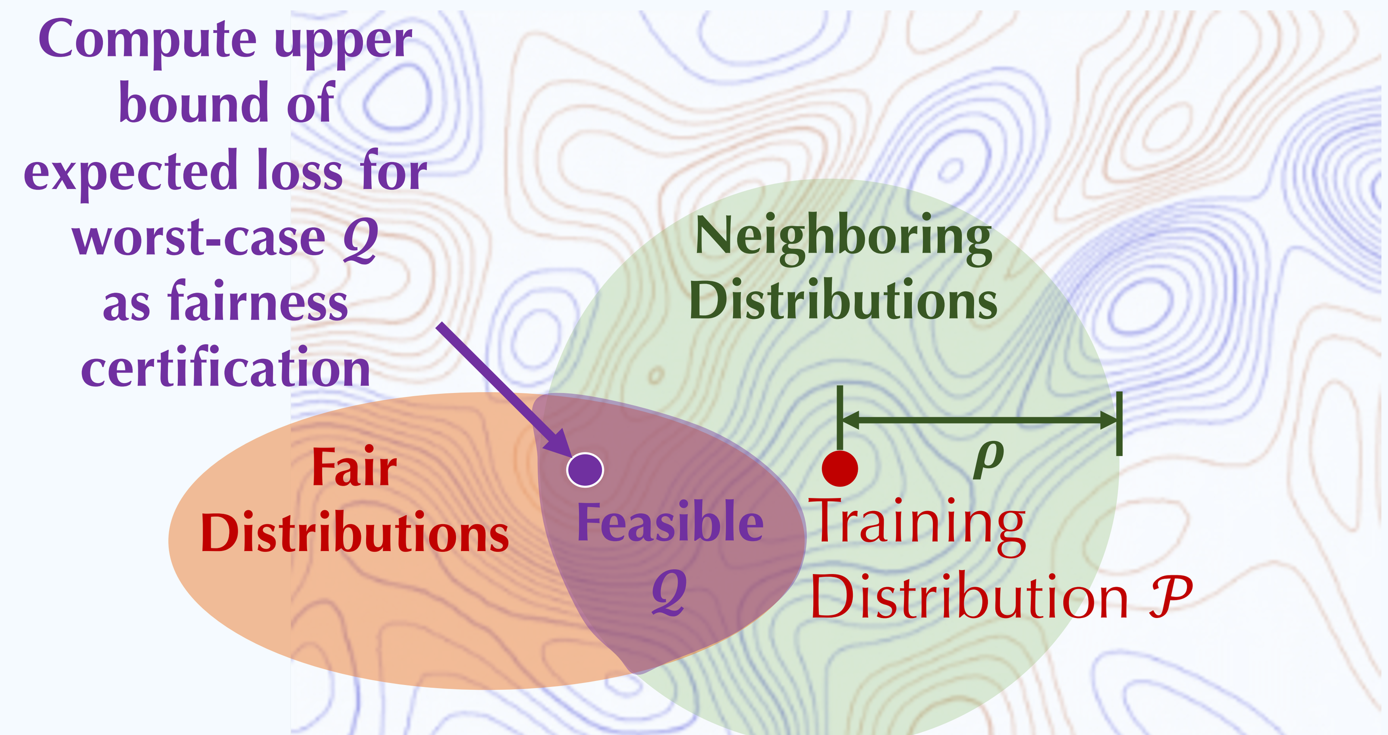
upper bounding $\max_{\mathcal{Q}} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)]$

s.t. $\mathrm{dist}(\mathcal{P},\mathcal{Q}) \le \rho$, $\mathcal{Q}$ is a fair distribution

- $\mathrm{dist}(\mathcal{P},\mathcal{Q}) \le \rho$ → **Test distribution $\mathcal{Q}$ not too far from training distribution**
- $\mathcal{Q}$ is a fair distribution → **Measure performance on distribution with fair base rate**

where

ρ	User-specified distance threshold
$\mathrm{dist}(\mathcal{P},\mathcal{Q})$	Hellinger distance between distributions



Loss landscape of $\mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)]$ w.r.t. $\mathcal{Q}$

Theoretical Observations

Distance Constraint	Decomposed to constraints on $\mathcal{P}_{s,y}$, $\mathcal{Q}_{s,y}$
Fair Distribution Constraint	Equal to constraints on $\Pr_{\mathcal{P}}[X_s = s, Y = y]$, $\Pr_{\mathcal{Q}}[X_s = s, Y = y]$

Distance constraint

$$\mathrm{dist}(\mathcal{P},\mathcal{Q}) \le \rho \Leftrightarrow$$

$$1 - \rho^2 - \sum_{s=1}^{S}\sum_{y=1}^{C} \sqrt{\Pr_{\mathcal{P}}[X_s = s, Y = y] \Pr_{\mathcal{Q}}[X_s = s, Y = y]} \left(1 - \mathrm{dist}(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y})^2\right) \le 0$$

Fair Distribution Constraint

- Consider discrete sensitive attribute X_s and label Y

Define fair distribution to be distribution with fair base rate:

$$\Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = s_a] = \Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = s_b], \forall y, s_a, s_b$$

➢ Sensitive attribute X_s has no effect on label Y at population level

- Such fair distribution admits unconstrained parameterization:

$$\Pr_{(X,Y)\sim\mathcal{Q}}[Y = y | X_s = s] = k_s r_y \quad (k_s, r_y \in [0,1])$$

Certification Procedure (Informal, Theorem 3)

Input: subpopulation statistics & subpopulation level constraints

1. Query subpopulation statistics:
$$\Pr_{\mathcal{P}}[X_s = s, Y = y], \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$$
2. Divide $k_s, r_y \in [0,1]$ into grids
3. In each grid:
   - **Known quantities**:
   $$\Pr_{\mathcal{P}}[X_s = s, Y = y], \Pr_{\mathcal{Q}}[X_s = s, Y = y], \mathbb{E}_{(X,Y)\sim\mathcal{P}_{s,y}}[\ell(h_\theta(X), Y)]$$
   - **Variables to optimize**: $\mathrm{dist}(\mathcal{P}_{s,y}, \mathcal{Q}_{s,y})^2$ (subject to distance constraints)
   - **Key variable to upper bound**: $\mathbb{E}_{(X,Y)\sim\mathcal{Q}_{s,y}}[\ell(h_\theta(X), Y)]$
   - ➢ Plug in Gramian bound [Weber et al, ICML 2022] to get upper bound
   - ➢ Optimize the upper bound with low-dimensional convex optimization
     - ➢ Bypass non-convexity with variable transforms
4. Maximization over all grids ⟹ **Output:** Certification of fairness!

Remarks:

- *For **sensitive shifting** setting (no distribution shift within each subpopulation, only portions among subpopulations shifted), we have **simpler** fairness certification procedure with **tighter** guarantees*
- *Framework **amenable to finite sampling error**: with high-confidence intervals of statistics, we provide high-confidence probabilistic certification.*
- *Framework **support any population loss function**, e.g., can bound group risk discrepancy*
- *Our fairness notion implies demographic parity (**DP**) and equalized odds (**EO**)*

Experimental Evaluation

Conclusions

Tightness: distance between gray points and black curve

➢ **Usually tight, especially in sensitive shifting setting**

Soundness: gray points always below black curve

➢ **Always sound**

More results & ablation studies in our paper!

x-axis: distance threshold ρ
y-axis: expected loss

For **sensitive shifting** setting:

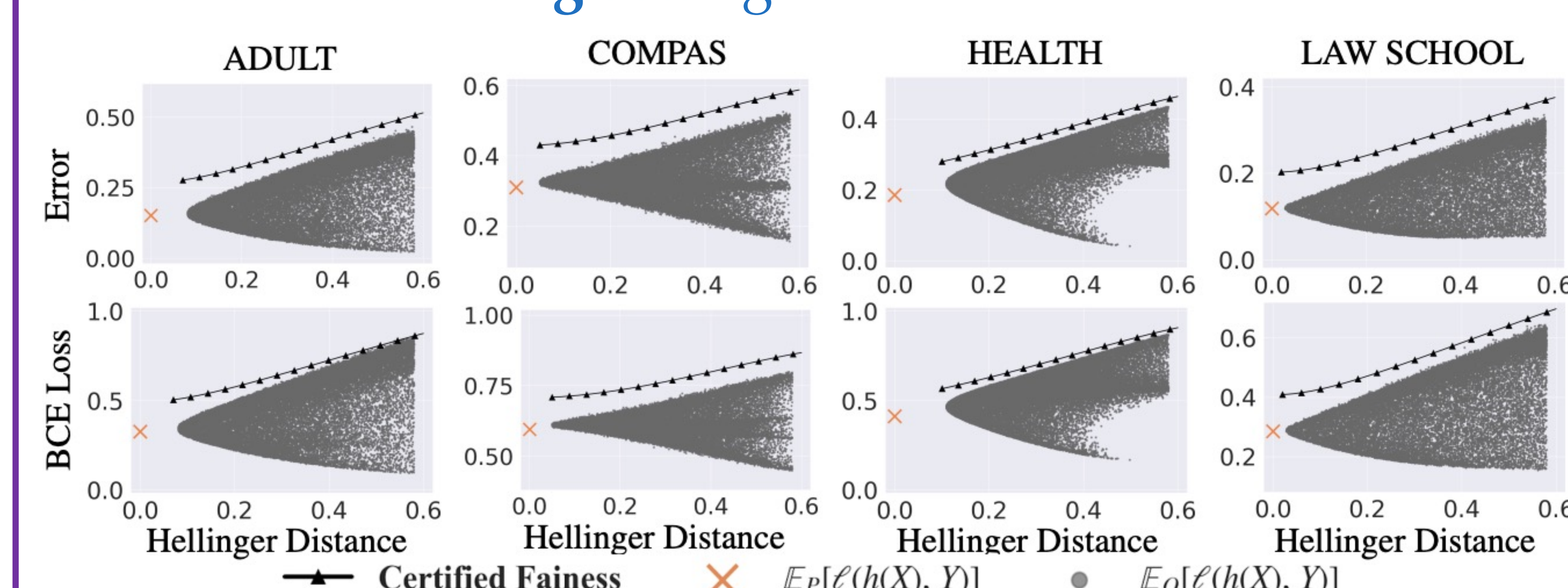


Figure 1: Certified fairness with sensitive shifting. Grey points are results on generated distributions ($\mathcal{Q}$) and the black line is our fairness certificate based on Thm. 2. We observe that our fairness certificate is usually tight.

For **general shifting** setting:

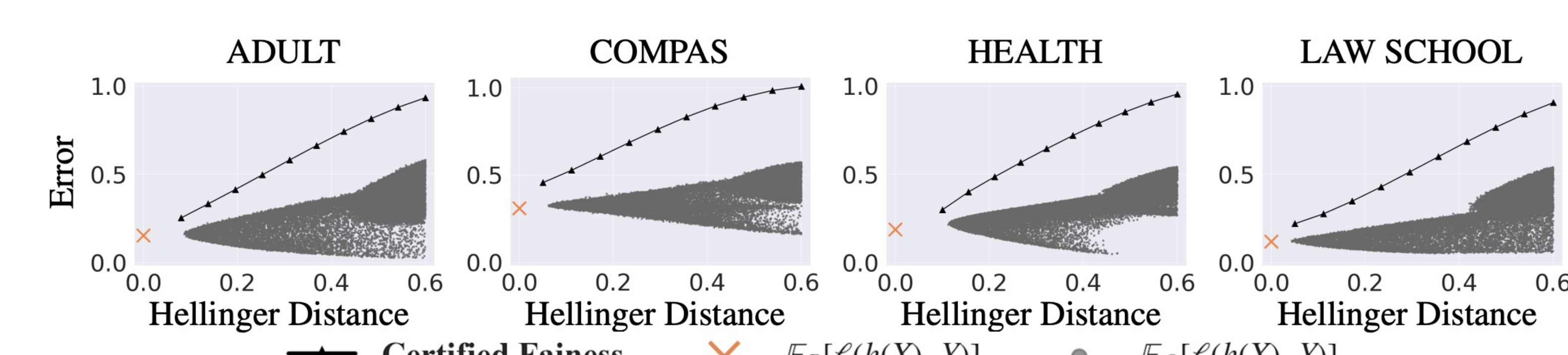


Figure 2: Certified fairness with general shifting. Grey points are results on generated distributions ($\mathcal{Q}$) and the black line is our fairness certificate based on Thm. 3. We observe that our fairness certificate is non-trivial.