Certifying Some Distributional Fairness with Subpopulation Decomposition

Mintong Kang*1, Linyi Li*1, Maurice Weber2, Yang Liu3, Ce Zhang2, Bo Li1

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

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 sensitive attribute \(X_s\) and label \(Y\)

\[
P = \sum_{s=1}^{S} \sum_{y=1}^{C} \Pr[X_s = s, Y = y] \cdot P_{s,y},
\]

\[
Q = \sum_{s=1}^{S} \Pr[X_s = s, Y = y] \cdot Q_{s,y}
\]

Problem Formulation

Given model \(h_0()\), compute an upper bound of its expected loss on a fair test distribution \(Q\), i.e.,

\[
\text{upper bounding } \max \mathbb{E}_{(X,Y) \sim Q}[f(h_0(X),Y)] \\text{s.t. } \text{dist}(P,Q) \leq \rho, \quad Q \text{ is a fair distribution}
\]

Certification Procedure (Informal, Theorem 3)

Input: subpopulation statistics & subpopulation level constraints

1. Query subpopulation statistics:
   \[
P_{r}[X_s = s, Y = y], \mathbb{E}_{(X,Y) \sim P_{s,y}}[f(h_0(X),Y)]
   \]
2. Divide \(k_s, r_s \in [0,1]\) into grids
3. In each grid:
   - \(P_s\)solves this certified fairness problem of an end ML model
   - \(Q\)solves this certified fairness problem of an end ML model

Remarks:
- For sensitive shifting setting (no distribution shift within each subpopulation, only portions among subpopulations shifted), we have simpler fairness certification 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)

Theoretical Observations

Fair Distribution Constraint
- Consider discrete sensitive attribute \(X_s\) and label \(Y\)
- Define fair distribution to be distribution with fair base rate:
  \[
  \Pr_{\phi_{Y|X}}[Y = y | X_s = s] = \Pr_{\phi_{Y|X}}[Y = y | X_s = s_b], \forall y, s_a, s_b
  \]
- Sensitive attribute \(X_s\) has no effect on label \(Y\) at population level
- Such fair distribution admits unconstrained parameterization:
  \[
  \Pr_{\phi_{Y|X}}[Y = y | X_s = s] = k_s r_y \quad (k_s, r_y \in [0,1])
  \]

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

Experimental Evaluation

For sensitive shifting setting:

- More results & ablation studies in our paper!

For general shifting setting:

- Figure 1: Certified fairness with sensitive shifting. Gray points are results on generated distributions (Q) and the black line is our fairness certificate based on Thm. 3. We observe that our fairness certificate is usually tight.

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