Input-agnostic Certified Group Fairness via Gaussian Parameter Smoothing

Jiayin Jin[1], Zeru Zhang[1], Yang Zhou[1], Lingfei Wu[2]

[1]Auburn University
[2]JD.COM Silicon Valley Research Center

ICML | 2022

Group Fairness

- Group fairness
  - A classifier makes accurate predictions and prevents itself from acting against specific groups
- Related work in empirical fairness
  - Follow manually-crafted heuristics to generate fair classifiers by solving non-convex problems
- Related work in certified fairness
  - Assume that the training data and the deployment data follow the same distribution

Problem Definition & Motivation

- Our goal
  - Certify the group fairness of classifiers with theoretical input-agnostic guarantees
  - No need to know the shift between training and deployment datasets w.r.t. sensitive attributes

- Motivation
  - Randomized data smoothing with the state-of-the-art certified robustness guarantees against worst-case attacks
  - Agnostic to input data and network architectures
  - Gaussian parameter smoothing for the certification of the input-agnostic group fairness

Classifier Smoothing

- Gaussian parameter smoothing

$$\hat{f}(x; W_k) = \mathop{\mathbb{E}}_{\Delta}\big(f(x; W_k + \Delta)\big), \text{ and}$$

$$\hat{f}(x; W) = \mathop{\mathbb{E}}_{\Delta}\big(f(x; W + \Delta)\big), \ \Delta \sim \mathcal{N}(0, \sigma^2 I)$$

- Training of optimal individual smooth classifiers

$$\hat{f}(x; W_k^*)$$

- Generation of overall smooth classifier $\hat{f}(x; W^*)$

$$W^* = \frac{W_1^* + \cdots + W_K^*}{K}$$

Theoretical Analysis

- Reformulate smooth classifiers as Nemytskii operator

$$\hat{N}(W)(\cdot) = \mathbb{E}\left(f(\cdot; W + \Delta)\right), \; \Delta \sim \mathcal{N}(0, \sigma^2 I)$$

- Input-agnostic global Lipschitz constant

$$\|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\| \leq \frac{\|W_1 - W_2\|_2}{\sqrt{2\pi}\sigma}$$

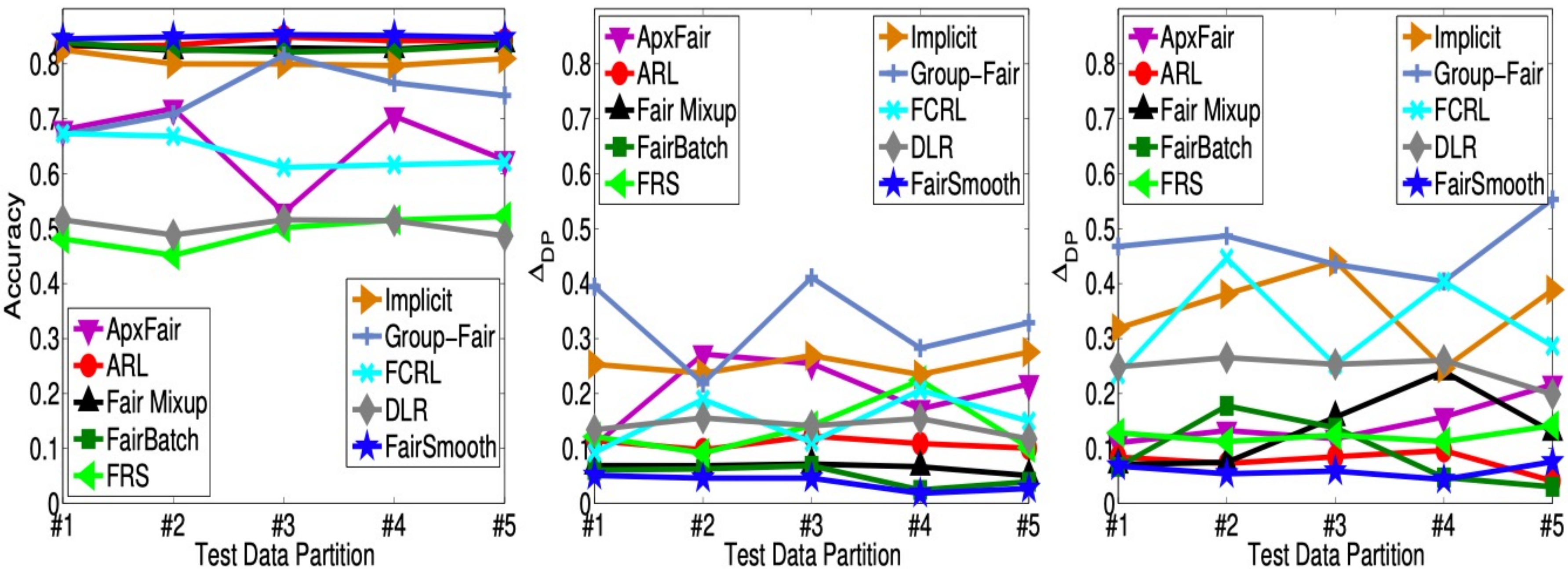
- Input-agnostic certified group fairness

$$|\Omega_k|^{-1} \|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^p(\Omega)} \leq \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$

if $1 \leq p < \infty$,

$$\|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^\infty(\Omega_k)} \leq \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$

Performance with Varying Test Data Partitions on Adult



(a) Accuracy (b) Δ_{DP} (c) Δ_{EO}