An Intersectional Definition of Fairness

Haeyoung Lee

April 15, 2025

Seoul National University

Why Intersectional Fairness?

- Existing fairness definitions often fail to protect minority or intersectional groups.
- **Intersectionality**: individuals may face multiple, overlapping sources of disadvantage (e.g., race and gender).
- We need fairness definitions that account for all combinations of protected attributes.

Baseline: Statistical Parity Subgroup Fairness (SF)

Definition I.1: A mechanism $M(x)$ is γ-statistical parity subgroup fair with respect to θ and a set G of group indicators $g : A \to \{0, 1\}$ if:

$$|P_{M,\theta}(M(x) = 1) - P_{M,\theta}(M(x) = 1 \mid g(s) = 1)| \cdot P_\theta(g(s) = 1) \leq \gamma \quad (1)$$

Notation:

- $x \in \chi$: input vector (e.g., an individual's features), $y \in \{0, 1\}$: binary output label
- $M(x)$: fair algorithm (e.g., a model that outputs y)
- $S_1, \ldots, S_p$: discrete protected attributes (e.g., race, gender, nationality)
- $A = S_1 \times S_2 \times \cdots \times S_p$: the Cartesian product of protected attribute spaces (i.e., all possible attribute combinations)
- $s \in A$: protected attribute tuple of an individual (e.g., (Black, Female))
- $g : A \to \{0, 1\}$: group indicator function, where $g(s) = 1$ means individual with s is in group g
- θ: data-generating distribution over input space χ
- $P_{M,\theta}$: model output probability under algorithm M and distribution θ
- γ: fairness tolerance parameter

Limitation: weights unfairness by group size $P_\theta(g(s) = 1)$, thus reducing the effect of minority groups.

Legal Motivation: The 80% Rule

- U.S. law provides the “80% rule” as a guideline for disparate impact.
- States that if the ratio of favorable outcomes between groups is less than 0.8, there is evidence of discrimination.
- Expressed mathematically as:

$$\frac{P(M(x) = 1 \mid \text{group A})}{P(M(x) = 1 \mid \text{group B})} < 0.8 \tag{2}$$

- The proposed definition, called Differential Fairness (DF), extends the 80% rule by introducing a tunable parameter ε, allowing for more flexible and continuous control over fairness across intersectional groups.

Proposed Definition: Differential Fairness (DF)

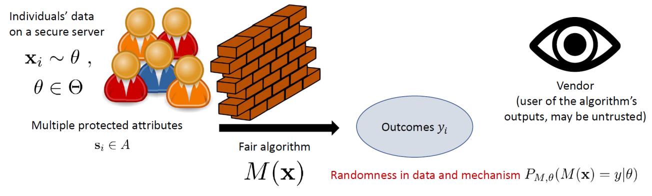


Fig. 1. Diagram of the setting for the proposed differential fairness criterion.

Figure 1: Diagram of the differential fairness setting.

Definition II.1: A mechanism $M(x)$ is ε-differentially fair with respect to (A, Θ) if:

$$e^{-\varepsilon} \leq \frac{P_{M,\theta}(M(x) = y \mid s_i)}{P_{M,\theta}(M(x) = y \mid s_j)} \leq e^{\varepsilon}, \quad \forall (s_i, s_j) \in A \times A, \ y \in \mathcal{Y} \quad (3)$$

Notation:

- $s_i, s_j \in A$: protected attribute tuples
- $\mathcal{Y}$: the range of possible output values of the mechanism $M(x)$
- ε: fairness parameter that bounds outcome probability ratios between groups
- Θ: a set of plausible data-generating distributions θ

Theoretical Guarantee: Intersectionality Property

Theorem IV.1 (Intersectionality Property): Let M be an ε-differentially fair mechanism in (A, Θ), where $A = S_1 \times S_2 \times \cdots \times S_p$, and let $D = S_a \times \cdots \times S_k$ be the Cartesian product of any nonempty proper subset of protected attributes in A. Then M is also ε-differentially fair in (D, Θ).

- Protecting intersectional groups *automatically* protects all subgroups.
- No need to separately enforce fairness at each attribute level.
- This provides a strong theoretical alignment with the goals of intersectionality.

Learning Fair Models under DF

Objective Function:

$$\min_{W} \left[L_X(W) + \lambda \cdot R_X(\varepsilon) \right] \tag{7}$$

Where:

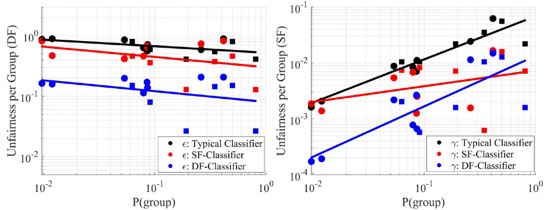
- W : model parameters of the classifier $M_W(x)$
- $L_X(W)$: prediction loss on data X (e.g., cross-entropy loss)
- λ : regularization coefficient to balance fairness and accuracy
- $R_X(\varepsilon) = \max(0, \varepsilon_{M_W(x)} - \varepsilon_1)$: fairness penalty

Notation:

- $\varepsilon_{M_W(x)}$: estimated DF violation level for the current model M_W
- ε_1: fairness threshold (e.g., 0 for strict DF)

Results

- **Dataset:** COMPAS (criminal recidivism prediction)
- **Protected attributes:** race and gender
- **Compared methods:** Typical Classifier (no fairness constraint), SF-Classifier (γ-Statistical Fairness), DF-Classifier (ε-Differential Fairness)
- **Metric:** Per-group unfairness vs. group probability (group size)



(a) Improvement in DF measures

(b) Improvement in SF measures

Fig. 2. Per-group measurements of (a) ϵ-DF and (b) γ-SF of the classifiers vs group size (probability), COMPAS dataset, calculated using Equations 1 and 3 with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. The methods improve fairness, both per group and overall, but SF-Classifier is seen to ignore minority groups in the overall γ-SF measurement, calculated as a worst-case over all groups.

Figure 2: Per-group measurements of (a) ε-DF and (b) γ-SF of the classifiers vs group size.

Result: DF-Classifier improves fairness for minority and intersectional groups better than SF-Classifier.

Thank you!