LoGoFair: Post-Processing for Local and Global Fairness in Federated Learning

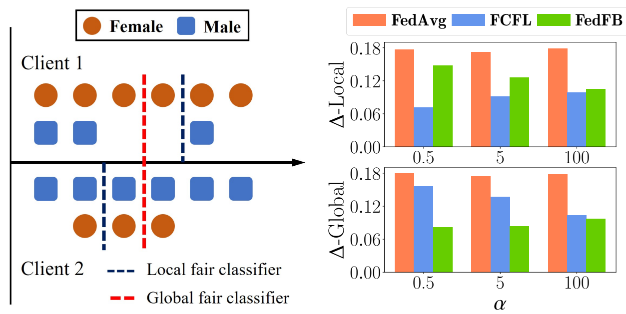
Li Zhang, Chaochao Chen*, Zhongxuan Han, Qiyong Zhong, Xiaolin Zheng

Zhejiang University

zhanglizl80@gmail.com, {zjuccc, zxhan, youngzhong, xlzheng}@zju.edu.cn

Abstract

Federated learning (FL) has garnered considerable interest for its capability to learn from decentralized data sources. Given the increasing application of FL in decision-making scenarios, addressing fairness issues across different sensitive groups (e.g., female, male) in FL is crucial. Current research often focuses on facilitating fairness at each client's data (*local fairness*) or within the entire dataset across all clients (*global fairness*). However, existing approaches that focus exclusively on either local or global fairness fail to address two key challenges: (**CH1**) *Under statistical heterogeneity, global fairness does not imply local fairness, and vice versa.* (**CH2**) *Achieving fairness under model-agnostic setting.* To tackle the aforementioned challenges, this paper proposes a novel post-processing framework for achieving both **Lo**cal and **Glo**bal **Fair**ness in the FL context, namely LoGoFair. *To address CH1*, LoGoFair endeavors to seek the Bayes optimal classifier under local and global fairness constraints, which strikes the optimal accuracy-fairness balance in the probabilistic sense. *To address CH2*, LoGoFair employs a model-agnostic federated post-processing procedure that enables clients to collaboratively optimize global fairness while ensuring local fairness, thereby achieving the optimal fair classifier within FL. Experimental results on three real-world datasets further illustrate the effectiveness of the proposed LoGoFair framework.



(a) Local/Global Fairness (b) Fairness under Heterogeneity

Figure 1: (a) Demonstrates a toy example of *local and global fairness* in the context of an one-dimensional classification problem. These fair classifiers ensure equal gender proportions in classification at local and global level (e.g. Local classifiers allocate 2/3 of samples from both genders to the left side for client 1). (b) Presents comparisons between local (FCFL) and global (FedFB) fair FL algorithms across varying levels of heterogeneity. A smaller α signifies more heterogeneity across clients, and a smaller Δ denotes a fairer model at local or global level.

1 Introduction

Federated learning is a distributed machine learning paradigm that enables multiple clients to collaboratively refine a shared model while preserving their data privacy (McMahan et al. 2017). With the growing integration of FL in high-stakes scenarios such as healthcare (Rieke et al. 2020; Chen et al. 2024a), finance (Chouldechova 2017a), and recommendation systems (Burke 2017), fairness is gaining prominence to prevent machine learning models from discriminating against any demographic group based on sensitive attributes, e.g. gender and race. Several methods exist to achieve group fairness in centralized settings (Agarwal et al. 2018; Alghamdi et al. 2022; Jovanović et al. 2023; Chen, Klochkov, and Liu 2024), these methods typically require direct access to entire datasets, thereby incurring high communication costs and privacy concerns if directly implemented in the FL environment.

To develop fairness guarantees for federated algorithms, this paper focuses on two key concepts of group fairness in FL: Local and Global Fairness (Cui et al. 2021; Ezzeldin et al. 2023; Hamman and Dutta 2024). **Local fairness** aims to develop models that deliver unbiased predictions across specific groups when evaluated on each client's local dataset. Since the models are ultimately deployed and applied in local environments, achieving local fairness is indispensable for promoting fair FL models. **Global fairness** focuses on identifying models that ensure similar treatment for sensitive groups within the entire dataset across all clients. In practice, models trained on large-scale aggregated datasets are inclined to learn inherent bias in data and exacerbate the treatment discrepancy of sensitive groups based as shortcuts to achieving high accuracy. (Geirhos et al. 2020; Chang and Shokri 2023). These global models typically fail to make impartial decisions and uphold societal fairness. Figure 1a

*Chaochao Chen is the corresponding author.

provides an example of local and global fairness, illustrating that these two fairness notions can differ. Therefore, both local and global fairness are crucial for achieving group fairness in the FL setting.

Recent works have introduced several techniques for achieving either local or global fairness. Most of them enhance fairness through dynamically adapting aggregation weights (Cui et al. 2021; Du et al. 2021a; Ezzeldin et al. 2023) or reweighting the training samples (Abay et al. 2020; Zeng, Chen, and Lee 2021). However, existing methods face certain challenges (CH) in achieving both local and global fairness. **CH1**: *Under statistical heterogeneity, local fairness does not imply global fairness, and vice versa.* (1) Statistical heterogeneity in FL leads to varying representations of sensitive groups across clients, causing local biases to differ from those in the globally aggregated dataset (Cui et al. 2021). As a result, methods that lessen global bias are unable to specifically tackle local bias and ensure fairness at the client level, similar to methods focused on local fairness. Previous work (Hamman and Dutta 2024) rigorously proved that neither local nor global fairness implies the other in FL, and further empirically examined this result across various cases of heterogeneity. (2) It is uncertain whether the local model will retain its debiasing capability after model aggregation, especially for nonlinear deep learning models (Chang and Shokri 2023). Figure 1b compares the fairness performance of methods tailored for either local or global fairness. The effectiveness of the comparison algorithms in the other fairness notions present a dramatic decline with increasing data heterogeneity. Hence, it is explicitly impracticable to achieve both local and global fairness by focusing solely on one of them within FL. **CH2:** *Achieving fairness under model-agnostic setting*. Most existing approaches employ the *in-processing* strategy (Cui et al. 2021; Du et al. 2021b; Ezzeldin et al. 2023), which entails intervening in the model training process and typically complicates the FL models. However, retraining the federated model is often impractical due to significant communication overhead and privacy concerns. An effective solution lies in investigating model-agnostic approaches to ensuring fairness in FL.

In this paper, we propose a novel federated post-processing framework for achieving both local and global fairness, namely LoGoFair. To the best of our knowledge, we are the first to offer both local and global fairness guarantees in the FL context. Generally, LoGoFair facilitates clients in collaboratively calibrating the pre-trained classifier under a specific procedure crafted to guarantee both fairness notions. To tackle **CH1**, we establish a specific characterization of the Bayesian optimal classifier under local and global fairness constraints. LoGoFair aims to achieve both fairness notions with minimal accuracy decline through learning this classifier. To tackle **CH2**, LoGoFair introduces a federated post-processing procedure that learns to calibrate federated models to the optimal fair classifier in a model-agnostic manner. This approach enables participating clients to collaboratively enhance global fairness while reinforcing their individual local fairness. The numerical results demonstrate that LoGoFair outperforms existing methods in achieving superior local and global fairness with competitive model accuracy. Moreover, experiments also illustrate that LoGoFair enables the flexible adjustment of the accuracy-fairness trade-off in the FL environment.

Our main contributions can be summarized as follows:

- We propose a novel post-processing framework named LoGoFair to achieve both local and global fairness in FL.
- We characterize an explicit formula of the Bayesian optimal classifier under local and global fairness constraints as the learning target of LoGoFair.
- We propose a federated post-processing procedure for clients to jointly optimize global fairness while guaranteeing local fairness, aiming to achieve the optimal fair classifier.
- Experimental results demonstrate that LoGoFair outperforms existing methods in achieving superior balances among accuracy, local fairness, and global fairness.

2 Related work

2.1 Group Fairness in Machine Learning

Group fairness in machine learning has grown rapidly over the last few years into a key area of trustworthy AI, especially in high-stakes decision-making systems, such as healthcare (Ahmad, Eckert, and Teredesai 2018; Caton and Haas 2024), criminal prediction (Berk et al. 2021) and recommendation systems (Li et al. 2021; Han et al. 2023, 2024a,b; Chen et al. 2024b). As summarized in previous work (Mehrabi et al. 2021), group fairness is broadly defined as the absence of prejudice or favoritism toward a sensitive group based on their inherent characteristics. In conventional centralized machine learning, common strategies for realizing group fairness can be classified into three categories: pre-processing, in-processing, and post-processing techniques. **Pre-processing** (Li and Liu 2022; Jovanović et al. 2023; Kang et al. 2020) approaches aim to modify training data to eradicate underlying bias before model training. **In-processing** (Kim et al. 2022; Li et al. 2023) methods are developed to achieve fairness requirements by intervention during the training process. **Post-processing** (Chzhen et al. 2020; Chen, Klochkov, and Liu 2024; Denis et al. 2024) methods adjust the prediction results generated by a given model to adapt to fairness constraints after the model is trained. In this paper, we propose a post-processing technique specifically tailored to ensure group fairness in FL environments.

2.2 Fair Federated Learning

Fairness faces new challenges in the FL context. Several recent works have introduced and investigated fairness concepts specific to FL, such as client-based fairness, and collaborative fairness. Nonetheless, the impact of FL on group fairness has yet to be thoroughly explored. Existing methods primarily utilize in-processing strategies to address global or local fairness issues.

Concerning local fairness, prior work (Chang and Shokri 2023) highlights potential detrimental effects of FL on the group fairness of individuals, while others (Cui et al. 2021) propose algorithms to enhance local fairness without compromising performance consistency. Concerning global fairness, two main approaches are adaptive reweighting techniques (Mohri, Sivek, and Suresh 2019; Ezzeldin et al. 2023; Zeng, Chen, and Lee 2021; Abay et al. 2020) and solving federated optimization objectives with relaxed differentiable fairness constraints or regularization (Du et al. 2021a; Wang et al. 2023; Dunda and Song 2024). (1) Reweighting techniques dynamically reweight clients or data during training time. (2) Optimization methods solve objectives with relaxed fairness constraints or regularization.

Besides, previous work (Hamman and Dutta 2024) theoretically investigated the interplay between local and global fairness in FL. They introduced AGLFOP to explore the theoretical limits of the accuracy-fairness trade-off, aiming to identify the optimal performance achievable given global data distributions and prediction outcomes. Challenges that persist include limited flexibility and generalization ability in addressing unfairness issues in prediction tasks within intricate FL scenarios.

3 Preliminaries

3.1 Fairness Notion

Let (X, A, Y) be a random tuple, where $X \in \mathcal{X}$ for some feature space $\mathcal{X} \in \mathbb{R}^d$, labels $Y \in \mathcal{Y} = \{0, 1\}$ for a binary classification problem, and the sensitive attribute $A \in \mathcal{A}$. In machine learning, the concept of group fairness aims to ensure that ML models provide equitable treatment to individuals with diverse sensitive attributes, such as gender, race, and age. Without loss of generality, this paper concentrates on two sensitive groups following (Zeng, Chen, and Lee 2021; Ezzeldin et al. 2023), with $\mathcal{A} = \{-1, 1\}$ representing the set of protected sensitive attributes. The goal of fair classification is to identify a attribute-aware classifier $h(x, a) : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ subject to the constraints imposed by the specified fairness criteria. In this paper, we mainly focus on two popular group fairness criteria:

- **Demographic Parity (DP)** (Dwork et al. 2012) emphasizes that the positive rate of predictor $\widehat{Y} = h(X, A)$ is equal in each sensitive group, measured by

$$\mathcal{M}_{DP} = \left| P(\widehat{Y} = 1|A = -1) - P(\widehat{Y} = 1|A = 1) \right|.$$

- **Equalized Odds (EO)** (Hardt, Price, and Srebro 2016) concentrates on equalizing the false positive and true positive rates in each sensitive group, measured by

$$\mathcal{M}_{EO} = \max_{y \in \{0,1\}} \Big| P(\widehat{Y} = 1|A = -1, Y = y) - P(\widehat{Y} = 1|A = 1, Y = y) \Big|.$$

3.2 Fairness in Federated Learning

A federated system consists of numerous decentralized clients, so that we consider the global data space as $\mathcal{S} = \{\mathcal{X}, \mathcal{A}, \mathcal{Y}, \mathcal{C}\}$, where $\mathcal{C}$ represents the client index set with $|\mathcal{C}|$ denoting the number of total clients. The global data distribution can be formally represented by joint random variable $S = (X, A, Y, C) \in \mathcal{S}$. FL focuses on the scenario in which the data is dispersed across $|\mathcal{C}|$ different clients, with each client c processing a local data dataset $\mathcal{D}_c$, $c \in [|\mathcal{C}|]$. Each sample in $\mathcal{D}_c$ is assumed to be drawn from local distribution, represented as $(x_{c,i}, a_{c,i}, y_{c,i})$, where $i \in [N_c]$, and N_c represents the number of samples for client c. Since local fairness act on clients' individual data distributions, we use a attribute-aware fair classifier $h(x, a, c)$ for each client to modify federated model. Here we introduce Bayes score function to characterize fairness notions $\eta(x, a) := P(Y = 1|X = x, A = a)$, which possesses a natural extension in FL:

$$\eta(x, a, c) := P(Y = 1 \mid X = x, A = a, C = c). \quad (1)$$

FL aims to minimize global risk $\mathcal{R}(h) := P(h(X, A, C) \neq Y)$ through minimizing the weighted average of the loss across all clients: $\min_\theta L(\theta) = \sum_{c=1}^{|\mathcal{C}|} a_c L_c(\theta)$, and the local objective $L_c(\theta) = \frac{1}{N_c} \sum_{(x,y) \in \mathcal{D}_c} \ell(f(x; \theta), y)$, where a_c signifies the importance coefficient; $\ell(\cdot, \cdot)$ denotes the loss function; f indicates the federated model.

Definition 1 (Local Fairness and Global Fairness) *The local fairness measures the disparity regarding sensitive groups aroused by the federated model when evaluated on the individual data distribution of each client, while the global fairness measures the disparity regarding sensitive groups on global data distribution across all clients.*

Formally, local data distribution can be represented by conditional distribution $P(X, A, Y|C)$, while global federated distribution is $P(X, A, Y)$. The disparity on the treatment of sensitive groups can be qualified by the fairness metrics ($\mathcal{M}_{DP}$ and $\mathcal{M}_{EO}$) introduced before.

4 LoGoFair

4.1 Overview

In this paper, we propose a novel post-processing framework to achieve both local and global fairness, namely LoGoFair. The goal of LoGoFair is to guarantee fairness in the FL setting with minimal accuracy sacrifice, while remaining compatible with either local or global fairness. Designed as a post-processing framework, LoGoFair can be conveniently adopted to mitigate bias in a wide range of existing FL algorithms. As illustrated in Figure 2, our proposed method comprises two primary phases in the post-processing framework. In the first phase, LoGoFair deduces the representation of the Bayes optimal federated fair classifier that realizes both local and global fairness with maximal accuracy from aggregated information. This classifier can be achieved by optimizing a convex minimization problem, as proved in Theorem 1. In the second phase, the optimization regarding the fair classifier is reformulated as a bilevel problem after the training stage, incorporating local fairness optimization and parameterized global fairness guidance shared across all clients. Inspired by this problem formulation, LoGoFair establishes a federated post-processing procedure that enables clients to collaboratively enhance global fairness while refining their local fairness. The next two sections will investigate these two phases in detail.

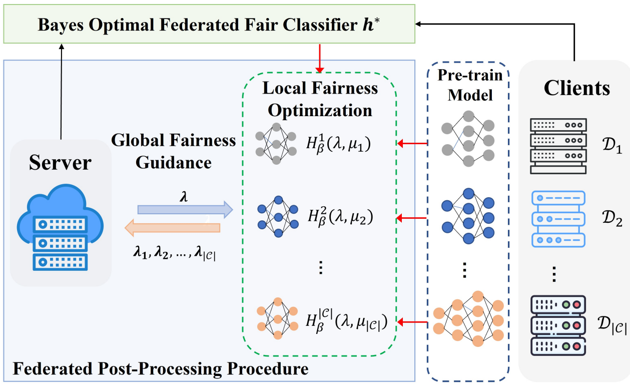


Figure 2: Overview of LoGoFair framework. The *Bayes optimal federated fair classifier*, which strikes the optimal accuracy-fairness balance, is identified as the objective. The *federated post-processing procedure* reformulates it as a bi-level problem incorporating local fairness optimization and global fairness guidance.

4.2 Bayes Optimal Federated Fair Classifier

As demonstrated in **CH1** and Figure 1b, models that focus solely on global or local fairness tend to fail in the other metric under statistical heterogeneity. This phenomenon compels us to account for fairness constraints at the both levels when developing fair FL models. To prevent excessive accuracy degradation due to fairness issues, we endeavor to solve the essential problem of what is the optimal classifier when local and global fairness restrictions are imposed.

We are interesting in either DP and EO metrics in this paper. For the sake of generality, we perform our analysis on *composite linear disparity*, a specific fairness metric that characterizes the group-wise disparity with Bayes optimal score. For a given distribution $\mathcal{P} \sim (X, A, Y)$, denote the marginal distribution of X by $\mathcal{P}^X$, the conditional distribution of $\mathcal{P}^X$ on sensitive group $A = a$ by $\mathcal{P}_a^X$. Given a classifier model $h(x, a)$, the *composite linear disparity* of h over $\mathcal{P}$ is defined as $\mathcal{M}_\phi(h) = \max_{k=1,\ldots,K} |\mathscr{D}_k(h)|$, and

$$\mathscr{D}_k(h) = \sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{X\sim\mathcal{P}_a^X} \left[\phi_k^a(\eta(X,a))h(X,a)\right].$$

where $\{\phi_k^a\}_{k=1}^K$ are linear functions that depend on a, and K is the number of linear disparities required to characterize the fairness criterion. This fairness metric encompasses DP and EO, along with other fairness notions such as Equality of Opportunity (Hardt, Price, and Srebro 2016) and Predictive Equality (Chouldechova 2017b). We give examples here to specify the fairness metrics represented by this notion.

Example 1 (DP and EO). *We have a sensitive attribute* $a \in \mathcal{A} = \{-1, 1\}$.
1) *Let* $K = 1$ *and* $\phi_1^a = a$, *we get DP metric.*
2) *Let* $K = 2$ *and* $\phi_1^a(w) = aw/P(Y = 1|A = a)$, $\phi_2^a(w) = a(1-w)/P(Y = 0|A = a)$, *we get EO metric.*

Furthermore, we extend *composite linear disparity* to the local and global fairness notions based on Definition 1. With the marginal distribution of X given sensitive attribute a and client c denoted by $\mathcal{P}_{a,c}^X$, the local and global fairness can be represented by

$$\begin{aligned}
\text{(Local)}:\quad & \mathcal{M}_\psi^{l,c}(h) = \max_{k_l=1,\ldots,K_{l,c}} \left|\mathscr{D}_{k_l}^{l,c}(h)\right|, \\
& \mathscr{D}_{k_l}^{l,c}(h) = \sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{X\sim\mathcal{P}_{a,c}^X} \left[\psi_{k_l}^{a,c}(\eta(X,a,c))h(X,a,c)\right], \\
\text{(Global)}:\quad & \mathcal{M}_\phi^{g}(h) = \max_{k_g=1,\ldots,K_g} \left|\mathscr{D}_{k_g}^{g}(h)\right|, \\
& \mathscr{D}_{k_g}^{g}(h) = \sum_{a\in\mathcal{A}} \sum_{c\in\mathcal{C}} \mathop{\mathbb{E}}_{X\sim\mathcal{P}_{a,c}^X} \left[\phi_{k_g}^{a,c}(\eta(X,a,c))h(X,a,c)\right],
\end{aligned}$$

where $\{\psi_{k_l}^{a,c}\}_{k_l=1}^{K_{l,c}}$ and $\{\phi_{k_g}^{a,c}\}_{k_g=1}^{K_g}$ are the linear functions determined by a and c, and $(K_{l,c}, K_g)$ are the number of local and global linear disparities. The detailed proof and more examples within FL are provided in **Appendix A**.

To investigate the optimal classifier with fairness guarantee within FL, we consider the situation that there is a unified fairness constraint in global level, and each client has additional local fairness restrictions in response to personalized demands. This problem can be interpreted as: *which classifier* h *minimizes the misclassification risk* $\mathcal{R}(h)$ *under the restriction that the local and global fairness measures* $(\mathcal{M}_\psi^{l,c}(h), \mathcal{M}_\phi^{g}(h))$ *are below given positive levels* (δ^l, δ^g). Denoting the set of all classifiers by $\mathcal{H}$, the problem can be formulated as

$$\begin{aligned}
& \min_{h\in\mathcal{H}} \mathcal{R}(h), \\
\text{s.t. } & |\mathscr{D}_{k_l}^{l,c}(h)| \le \delta^{l,c}, c \in [C], k_l \in [K_{l,c}], \\
& |\mathscr{D}_{k_g}^{g}(h)| \le \delta^{g}, k_g \in [K_g].
\end{aligned} \tag{2}$$

The solution of Problem (2) is identified as the *Bayes optimal federated fair classifier*, which theoretically achieves the optimal trade-off among accuracy, global fairness and local fairness in the FL context.

In order to establishing the closed form of optimal fair classifier, we introduce the following mild assumption.

Assumption 1 (η-continuity). *For each client* c, *we assume that the mappings* $t \mapsto P(\eta(X, A, C) \le t \mid A = a, C = c)$ *are almost surely continuous restricted to* $[0, 1]$.

Assumption 1 implies that the conditional probability function, when regarded as a random variable, has no atoms. It can be fulfilled from any estimator $\eta(x, a, c)$ and conditional distribution $\mathcal{P}_{a,c}^X$ by adding slight random noises (Chzhen et al. 2019, 2020; Denis et al. 2024). In brief, this assumption is of little or no concern in practice, as demonstrated in our experiments in section 5. In the following result, we verify that the optimal federated fair classifier can be attained by calibrating the Bayes optimal score function.

Theorem 1 (Beyes optimal federated fair classifier). *Under Assumption 1, for* $p_{a,c} := P(A = a, C = c)$, $\phi^{a,c} := [\phi_{k_g}^{a,c}]_{k_g=1}^{K_g}$ *and* $\psi^{a,c} := [\phi_{k_l}^{a,c}]_{k_l=1}^{K_{l,c}}$, h^* *is the Beyes optimal federated fair classifier which solves problem* (2) *if*

$$h^*(x, a, c) = \mathbb{I}\left[F(\lambda^*, \mu^*, x, a, c) \ge 0\right],$$

and calibration function

$$
\begin{aligned}
&F(\lambda^*, \mu^*, x, a, c) \\
&\quad = p_{a,c}(2\eta(x,a,c) - 1) - (\lambda_1^* - \lambda_2^*)^T \phi^{a,c}(\eta(x,a,c)) \\
&\qquad - (\mu_{1,c}^* - \mu_{2,c}^*)^T \psi^{a,c}(\eta(x,a,c)).
\end{aligned} \tag{3}
$$

where $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^{2K_g}$, $\mu = [\mu_c]_{c=1}^{|\mathcal{C}|}$, $\mu_c = (\mu_{1,c}, \mu_{2,c}) \in \mathbb{R}^{2K_{l,c}}$. Parameters λ^*, μ^* are determined from the convex minimization problem $(\lambda^*, \mu^*) \in \arg\min_{\lambda,\mu \geq 0} H(\lambda, \mu)$,

$$
\begin{aligned}
H(\lambda, \mu) = &\sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{X \sim \mathcal{P}_{a,c}^X} [(F(\lambda, \mu, X, a, c))_+] \\
&+ \delta^g (\lambda_1 + \lambda_2)^T \mathbf{1}_{K_g} + \delta^{l,c} \sum_{c \in \mathcal{C}} (\mu_{1,c} + \mu_{2,c})^T \mathbf{1}_{K_{l,c}},
\end{aligned} \tag{4}
$$

where $(\cdot)_+$ refers to $\max(\cdot, 0)$, $\mathbf{1}_K$ is the ones vector with dimension K. The optimality of λ^* provides global fairness guarantee and the optimality of μ_c^* provides local fairness guarantee for client c. (Proof see ***Appendix B****.)*

Remark This theorem can be applied to the situation when we focus solely on local or global fairness. The only adjustment is to simplify problem (2) by removing unnecessary local or global fairness constraints.

4.3 Federated Post-Processing Procedure

To develop model-agnostic fair FL approaches (**CH2**) while preserving data privacy, our characterization of optimal fair classifier naturally suggest a post-processing algorithm that estimates unknown Bayes optimal score η with pre-trained probabilistic classifier. Therefore, LoGoFair proposes a federated post-processing procedure to solve the optimal fair classifier, facilitating the collaborative enhancement of both local and global fairness among clients, as illustrated in Figure 2. In this framework, client-specific parameter μ_c guarantees local fairness, while λ serves as the global fairness guidance, according to Theorem 1.

Firstly, we investigate the estimation of the optimal fair classifier in Theorem 1 and formulate (4) as a bi-level problem. For Bayes optimal score η, while it is not typically possible to identify these ground-truth Bayes optimal score functions in practice, data-driven learning procedure allows us to train probabilistic classifiers as empirical estimators. We can utilize a wide range of FL approaches to obtain precisely probabilistic classifier trained using the cross-entropy loss (McMahan et al. 2017; Tan et al. 2022). For the purpose of further mitigating the estimation error in probability scores generated by the FL model,we adopt model calibration (Guo et al. 2017) to calibrate the learned FL classifier.

To estimate (λ^*, μ^*), notice that the formula of $H(\lambda, \mu)$ involves local marginal distributions $\mathcal{P}_{a,c}^X$ and some statistics in $\phi^{a,c}, \psi^{a,c}$ which are necessary for evaluating fairness. Therefore, we propose to calibrate fairness via post-processing in a validation dataset $\mathcal{D}^{val} = \bigcup_{c \in \mathcal{C}} \mathcal{D}_c^{val}$. Denoting the validation data of sensitive group a in client c as $\mathcal{D}_{a,c}^{val} = \{x_i^{a,c}\}_{i=1}^{N_{a,c}}$, the empirical estimation of (3) can be represented as

$$
\begin{aligned}
&\widehat{F}(\lambda, \mu, x, a, c) \\
&\quad = \widehat{p}_{a,c}(2\widehat{\eta}(x,a,c) - 1) - (\lambda_1 - \lambda_2)^T \widehat{\phi}^{a,c}(\widehat{\eta}(x,a,c)) \\
&\qquad - (\mu_{1,c}^* - \mu_{2,c}^*)^T \widehat{\psi}^{a,c}(\widehat{\eta}(x,a,c)).
\end{aligned} \tag{5}
$$

Furthermore, $\min_{\lambda,\mu \geq 0} H(\lambda, \mu)$ in (4) can be formulated as a bi-level problem with estimator $\widehat{F}$,

$$
\min_{\lambda \geq 0} \left\{ \widehat{H}(\lambda) = \sum_{c \in \mathcal{C}} \widehat{H}^c(\lambda) \right\}, \quad \widehat{H}^c(\lambda) := \min_{\mu_c \geq 0} \widehat{H}^c(\lambda, \mu_c), \tag{6}
$$

where the local fairness optimization task is

$$
\begin{aligned}
\widehat{H}^c(\lambda, \mu_c) = &\sum_{a \in \mathcal{A}} \frac{1}{N_{a,c}^{val}} \sum_{i=1}^{N_{a,c}^{val}} \left(\widehat{F}(\lambda, \mu, x_i^{a,c}, a, c) \right)_+ \\
&+ \frac{\delta^g}{|\mathcal{C}|} (\lambda_1 + \lambda_2)^T 1_{K_g} + \delta^{l,c} (\mu_{1,c} + \mu_{2,c})^T 1_{K_{l,c}}.
\end{aligned} \tag{7}
$$

Secondly, LoGoFair introduce a federated optimization algorithm tailored to solve the bi-level optimization (6). It is clear that $\widehat{H}(\lambda)$ and $\widehat{H}^c(\lambda, \mu_c)$ are still convex with respect to λ and μ_c but not smooth. Clients can locally apply subgradient descent (Bubeck 2015) or grid search (Chen, Klochkov, and Liu 2024) to enforce local fairness. However, considering that data is stored on the client side and must remain confidential, we are unable to optimize λ directly. Although it is possible to use subgradient in place of the gradient in traditional federated framework, such an approach is usually ineffective owing to the intrinsic slow convergence (Yuan, Zaheer, and Reddi 2021). To make the objective function smooth enough, we use the logarithmic exponential relaxation to replace the operator $(\cdot)_+$ in (7),

$$
r_\beta(x) = \frac{1}{\beta} \log(1 + \exp(\beta x)).
$$

Whenever $\beta \to \infty$, $r_\beta(x)$ reduces to $\max(x, 0)$.

To solve the federated post-processing problem (6), projection gradient descent is utilized to approach the optimal solution. We present the federated post-processing procedure in Algorithm 1 of **Appendix C**, along with its **efficiency**, **privacy** analysis, and additional discussion.

5 Experiment

In this section, we comprehensive evaluate the proposed LoGoFair method on three publicly available real-world datasets. Here we conduct extensive experiments to answer the following Research Questions (RQ): **RQ1**: Does LoGoFair outperform the existing methods in effectively achieving a balance between accuracy and fairness in FL? **RQ2**: Is LoGoFair capable of adjusting the trade-off between accuracy and local-global fairness (sensitivity analysis)? **RQ3**: How do client number influence the performance of LoGoFair? **RQ4**: How about the efficiency of LoGoFair?

5.1 Datasets and Experimental Settings

Due to space limitations, the detailed information in this section is provided in **Appendix D.1**.

α	Dataset	Adult			ENEM			CelebA		
	Method	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)
0.5	FedAvg	0.8381	0.1922	0.1759	0.7266	0.2030	0.1953	0.8934	0.1308	0.1441
	FedFB	0.8158	0.1174	0.0767	0.7089	0.1625	0.0801	0.8705	0.1064	0.0723
	FairFed	0.8079	0.1416	0.0956	0.6974	0.1721	0.1002	0.8555	0.1163	0.0975
	FCFL	0.8167	0.0832	0.1479	0.6997	0.0900	0.1837	0.8391	0.0827	0.1373
	LoGoFair$_g$	**0.8218**	**0.0889**	**0.0183***	**0.7111**	**0.0574**	**0.0141***	**0.8698**	**0.0581**	**0.0154***
	LoGoFair$_l$	**0.8252***	**0.0367***	**0.0468**	**0.7129***	**0.0241***	**0.0473**	**0.8734***	**0.0196***	**0.0515**
	LoGoFair$_{l\&g}$	**0.8214**	**0.0489**	**0.0204**	**0.7109**	**0.0281**	**0.0151**	**0.8710**	**0.0312**	**0.0295**
5	FedAvg	0.8418	0.1820	0.1725	0.7268	0.1894	0.1814	0.8962	0.1394	0.1461
	FedFB	0.8243	0.1134	0.0658	0.7116*	0.1502	0.0807	0.8677	0.0856	0.0532
	FairFed	0.8157	0.1264	0.0971	0.7067	0.1600	0.1173	0.8516	0.1045	0.0874
	FCFL	0.8134	0.0673	0.1297	0.7031	0.0623	0.1073	0.8699	0.0585	0.0941
	LoGoFair$_g$	**0.8264***	**0.0399**	**0.0104***	**0.7107**	**0.0548**	**0.0112***	**0.8721**	**0.0484**	**0.0192***
	LoGoFair$_l$	**0.8237**	**0.0252***	**0.0215**	**0.7065**	**0.0133***	**0.0423**	**0.8730***	**0.0267***	**0.0475**
	LoGoFair$_{l\&g}$	**0.8244**	**0.0259**	**0.0138**	**0.7049**	**0.0209**	**0.0142**	**0.8704**	**0.0279**	**0.0242**
100	FedAvg	0.8466	0.1802	0.1759	0.7279	0.1852	0.1746	0.8997	0.1352	0.1467
	FedFB	0.8197	0.0901	0.0790	0.7047	0.0821	0.0698	0.8715	0.0736	0.0698
	FairFed	0.8267	0.0977	0.1086	0.7072	0.1027	0.0962	0.8617	0.0950	0.0862
	FCFL	0.8194	0.0613	0.0897	0.7014	0.0633	0.0871	0.8659	0.0587	0.0673
	LoGoFair$_g$	**0.8297***	**0.0368**	**0.0282***	**0.7118***	**0.0228**	**0.0114***	**0.8724***	**0.0265**	**0.0182***
	LoGoFair$_l$	**0.8288**	**0.0335***	**0.0379**	**0.7107**	**0.0189***	**0.0209**	**0.8708**	**0.0175***	**0.0213**
	LoGoFair$_{l\&g}$	**0.8283**	**0.0362**	**0.0297**	**0.7061**	**0.0193**	**0.0119**	**0.8712**	**0.0211**	**0.0295**

* The bold text indicates the result of LoGoFair. The best results are marked with *. The second-best results are underlined.
* All outcomes pass the significance test, with a p-value below the significance threshold of 0.05.
* We use **FedAvg** as the baseline for optimal accuracy, without comparing it in terms of the accuracy-fairness trade-off.

Table 1: Comparison experimental result.

Datasets We consider three real-world benchmarks, **Adult** (Asuncion, Newman et al. 2007), **ENEM** (INEP 2018), and **CelebA** (Zhang et al. 2020), which are well-established for assessing fairness issues in FL (Ezzeldin et al. 2023; Chang and Shokri 2023; Duan et al. 2024).

Baselines Since *no previous work was found that endeavors to simultaneously achieve local and global fairness* within a FL framework, we compare the performance of LoGoFair with **FedAvg** (McMahan et al. 2017) and three SOTA methods tailored for either global or local fairness, namely **FairFed** (Ezzeldin et al. 2023), **FedFB** (Zeng, Chen, and Lee 2021), **FCFL** (Cui et al. 2021). Meanwhile, we adapt LoGoFair to focus solely on local or global fairness in FL, denoted as LoGoFair$_l$ and LoGoFair$_g$. LoGoFair$_{l\&g}$ indicates the algorithm simultaneously achieving local and global fairness.

Evaluation Protocols *(1) Firstly*, we partition each dataset into a 70% training set and the remaining 30% for test set, while post-processing models use half of training set as validation set following previous post-processing works (Xian, Yin, and Zhao 2023; Chen, Klochkov, and Liu 2024). *(2) Secondly*, to simulate the statistical heterogeneity in FL context, we control the heterogeneity of the sensitive attribute distribution at each client by determining the proportion of local sensitive group data based on a Dirichlet distribution $Dir(\alpha)$ as proposed by Ezzeldin et al. (2023). In this case, each client will possess a dominant sensitive group, and a smaller value of α will further reduce the data proportion of the other group, which *indicates greater heterogeneity across clients*. *(3) Thirdly*, The number of participating clients is set to 5 to simulate the FL environment. *(4) fourthly*, we evaluate the FL model with Accuracy (Acc), global fairness metric $\mathcal{M}^{global}$ and maximal local fairness metric among clients $\mathcal{M}^{local}$. Since we are interesting in DP and EO criteria, the model's fairness is assessed by local and global DP/EO metrics ($\mathcal{M}_{DP/EO}^{local}, \mathcal{M}_{DP/EO}^{global}$), smaller values of fairness metrics denote a fairer model.

5.2 Overall Comparison (RQ1)

We conduct extensive experiments to compare LoGoFair with other existing fair FL baseline under varying degrees of statistical heterogeneity, and the results of DP criterion ($\mathcal{M}_{DP}$) are presented in Table 1. Here we set $\delta^{l,c} = \delta^g = 0.01$. It is essential to note that obtaining both high accuracy and fairness is challenging due to the inherent trade-off between these metrics. All the results of EO criterion ($\mathcal{M}_{EO}$) are reported in **Appendix D.3**.

Comparison result on various datasets. From the table we can see that LoGoFair generally demonstrates superior performance in achieving a balance between accuracy and fairness. Compared with FedFB and FairFed, which enables to achieve global fairness, LoGoFair gets both better accuracy and global fairness. In terms of local fairness, LoGoFair also surpasses FCFL in accuracy and fairness behavior. This results strongly confirm the effectiveness of our approach.

Ablation study. To demonstrate that LoGoFair is also applicable to scenarios focusing exclusively on either local or global fairness, we introduced LoGoFair$_l$ and LoGoFair$_g$. The results indicate that, with a single fairness constraint, these methods typically perform better in the targeted fairness notion compared to LoGoFair$_{l\&g}$. For local fairness, LoGoFair$_l$ implement fairer federated model in local level with best accuracy in most cases compared to other methods. Similarly, LoGoFair$_g$ also outperforms all other methods in

Dataset	Adult			ENEM			CelebA		
(δ_l, δ_g)	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)	Acc (↑)	$\mathcal{M}_{DP}^{local}$ (↓)	$\mathcal{M}_{DP}^{global}$ (↓)
$(0.00, 0.00)$	0.8186	0.0439	0.0035	0.7087	0.0187	0.0005	0.8607	0.0089	0.0015
$(0.02, 0.00)$	0.8209	0.0488	0.0040	0.7093	0.0243	0.0009	0.8688	0.0243	0.0039
$(0.04, 0.00)$	0.8209	0.0543	0.0062	0.7104	0.0459	0.0012	0.8701	0.0459	0.0052
$(0.00, 0.02)$	0.8235	0.0454	0.0339	0.7113	0.0192	0.0215	0.8713	0.0192	0.0215
$(0.02, 0.02)$	0.8255	0.0507	0.0366	0.7132	0.0261	0.0314	0.8732	0.0361	0.0314
$(0.04, 0.02)$	0.8243	0.0634	0.0400	0.7140	0.0492	0.0367	0.8740	0.0462	0.0367
$(0.00, 0.04)$	0.8238	0.0423	0.0352	0.7115	0.0220	0.0238	0.8715	0.0242	0.0438
$(0.02, 0.04)$	0.8252	0.0548	0.0448	0.7136	0.0332	0.0379	0.8736	0.0432	0.0479
$(0.04, 0.04)$	0.8265	0.0667	0.0516	0.7157	0.0540	0.0515	0.8757	0.0604	0.0505

Table 2: Accuracy-Fairness Trade-off (Sensitivity Analysis).

balancing accuracy and global fairness. Meanwhile, notice that LoGoFair$_{l\&g}$ generally provides both local and global fairness guarantee at the cost of slight accuracy degradation compared to LoGoFair$_l$ and LoGoFair$_g$.

Impact of statistical heterogeneity. The most significant cause of the inconsistency between local and global fairness is statistical heterogeneity in FL. Table 1 explicitly presents that the gap between local fairness and global fairness progressively narrows, as the variation in client data distributions lessens. Our proposed method still outperforms other approaches in most case under varying degrees of heterogeneity, further demonstrating its robustness to statistical heterogeneity in FL environments.

5.3 Accuracy-Fairness Trade-Off (RQ2)

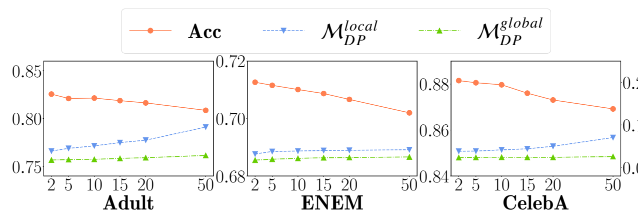
To investigate the capability of LoGoFair in adjusting accuracy-fairness trade-off, we report the Acc, $\mathcal{M}_{DP}^{local}$ and $\mathcal{M}_{DP}^{global}$ under different fairness relaxation of (δ^l, δ^g) with $\alpha = 0.5$ on all three datasets in Table 2. Here we also set the local fairness relaxation $\delta^{l,c}$ for each client to the same value, denoted as δ^l.

Sensitivity analysis. It is important to assess the sensitivity of LoGoFair's performance with respect to sensitivity (δ^l, δ^g). Table 2 reveals that with a constant global constraints δ^g, an decreasing in δ^l results in lower Acc and local fairness, indicating that while the model achieves superior local fairness, its overall performance declines. Similarly, holding δ^l constant allows the control of global fairness through adjustments in δ^g. These results confirm our claim that LoGoFair can flexibly adjust the accuracy-fairness trade-off within FL.

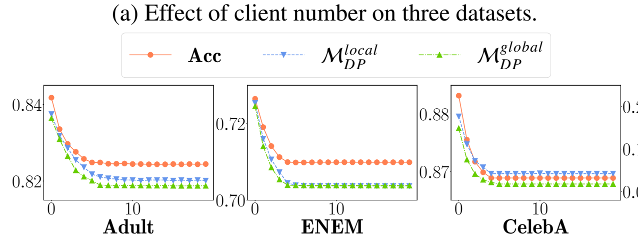
5.4 Scalability Concerning Client Number (RQ3)

We examine the performance of LoGoFair across a range of 2 to 50 clients on all three datasets using heterogeneity level $\alpha = 0.5$. The results are shown in Figure 3a.

Scalability. Local fairness metric in the Adult and CelebA datasets presents a minor increase with the growing number of clients. This trend can be attributed to the reduction in data samples available for local fairness evaluation, which exacerbates the estimation error of federated post-processing procedure. Other measures reveals slight fluctuations in accuracy and fairness metrics, underscoring the model's robustness to variations in client number.



(a) Effect of client number on three datasets.



(b) Effect of communication rounds on three datasets.

Figure 3: Scalability and effectiveness analysis.

5.5 Effectiveness Analysis (RQ4)

In this section, we conduct out experiments to examine the communication cost introduced by LoGoFair.

Communication Efficiency. The behavior of LoGoFair is monitored for different values of communication rounds T, leading to the results shown in Figure 3b. The rapid convergence of the three metrics to stable values within 10 rounds across three datasets empirically confirms the effectiveness.

6 Conclusion

This paper proposes a novel post-processing framework for achieving **Lo**cal and **Glo**bal **Fair**ness within FL, namely LoGoFair. To the best of our knowledge, we are the first to offer local and global fairness in FL. The goal of LoGoFair is to learn the Bayes optimal classifier under local and global fairness constraints in order to achieve both fairness notions with maximal accuracy. Furthermore, LoGoFair introduces the federated post-processing procedure to solve the optimal fair classifier in a model-agnostic manner accounting for communication cost and data privacy. This approach enables participating clients to collaboratively enhance global fairness while refining their local fairness. Experiments on three publicly available real-world datasets confirm that LoGoFair outperforms existing methods in achieving superior local and global fairness with competitive accuracy.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62172362).

References

Abay, A.; Zhou, Y.; Baracaldo, N.; Rajamoni, S.; Chuba, E.; and Ludwig, H. 2020. Mitigating Bias in Federated Learning. arXiv:2012.02447.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Ahmad, M. A.; Eckert, C.; and Teredesai, A. 2018. Interpretable Machine Learning in Healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, 559–560. New York, NY, USA: Association for Computing Machinery. ISBN 9781450357944.

Alghamdi, W.; Hsu, H.; Jeong, H.; Wang, H.; Michalak, P.; Asoodeh, S.; and Calmon, F. 2022. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35: 38747–38760.

Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1): 3–44.

Bubeck, S. 2015. Convex Optimization: Algorithms and Complexity. arXiv:1405.4980.

Burke, R. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

Chang, H.; and Shokri, R. 2023. Bias propagation in federated learning. *arXiv preprint arXiv:2309.02160*.

Chen, C.; Feng, X.; Li, Y.; Lyu, L.; Zhou, J.; Zheng, X.; and Yin, J. 2024a. Integration of large language models and federated learning. *Patterns*, 5(12).

Chen, C.; Zhang, J.; Zhang, Y.; Zhang, L.; Lyu, L.; Li, Y.; Gong, B.; and Yan, C. 2024b. CURE4Rec: A Benchmark for Recommendation Unlearning with Deeper Influence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chen, W.; Klochkov, Y.; and Liu, Y. 2024. Post-hoc bias scoring is optimal for fair classification. In *The Twelfth International Conference on Learning Representations*.

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

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

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33: 19137–19148.

Cui, S.; Pan, W.; Liang, J.; Zhang, C.; and Wang, F. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34: 26091–26102.

Denis, C.; Elie, R.; Hebiri, M.; and Hu, F. 2024. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130): 1–46.

Du, W.; Xu, D.; Wu, X.; and Tong, H. 2021a. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 181–189. SIAM.

Du, W.; Xu, D.; Wu, X.; and Tong, H. 2021b. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 181–189. SIAM.

Duan, Y.; Tian, Y.; Chawla, N.; and Lemmon, M. 2024. Post-Fair Federated Learning: Achieving Group and Community Fairness in Federated Learning via Post-processing. *arXiv preprint arXiv:2405.17782*.

Dunda, G. W. M.; and Song, S. 2024. Fairness-aware Federated Minimax Optimization with Convergence Guarantee. arXiv:2307.04417.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ezzeldin, Y. H.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, A. S. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 7494–7502.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hamman, F.; and Dutta, S. 2024. Demystifying Local & Global Fairness Trade-offs in Federated Learning Using Partial Information Decomposition. In *The Twelfth International Conference on Learning Representations*.

Han, Z.; Chen, C.; Zheng, X.; Li, M.; Liu, W.; Yao, B.; Li, Y.; and Yin, J. 2024a. Intra- and Inter-group Optimal Transport for User-Oriented Fairness in Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8): 8463–8471.

Han, Z.; Chen, C.; Zheng, X.; Liu, W.; Wang, J.; Cheng, W.; and Li, Y. 2023. In-processing User Constrained Dominant Sets for User-Oriented Fairness in Recommender Systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 6190–6201. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Han, Z.; Chen, C.; Zheng, X.; Zhang, L.; and Li, Y. 2024b. Hypergraph Convolutional Network for User-Oriented Fairness in Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 903–913. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

INEP, C. D. É. 2018. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Boletim de Serviço Eletrônico em*, 30: 04.

Jovanović, N.; Balunovic, M.; Dimitrov, D. I.; and Vechev, M. 2023. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, 15401–15420. PMLR.

Jovanović, N.; Balunovic, M.; Dimitrov, D. I.; and Vechev, M. 2023. FARE: Provably Fair Representation Learning with Practical Certificates. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 15401–15420. PMLR.

Kang, J.; He, J.; Maciejewski, R.; and Tong, H. 2020. InFoRM: Individual Fairness on Graph Mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 379–389. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Kim, D.; Kim, K.; Kong, I.; Ohn, I.; and Kim, Y. 2022. Learning fair representation with a parametric integral probability metric. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 11074–11101. PMLR.

Li, P.; and Liu, H. 2022. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12917–12930. PMLR.

Li, T.; Guo, Q.; Liu, A.; Du, M.; Li, Z.; and Liu, Y. 2023. FAIRER: Fairness as Decision Rationale Alignment. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 19471–19489. PMLR.

Li, Y.; Chen, H.; Fu, Z.; Ge, Y.; and Zhang, Y. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021*, WWW '21, 624–632. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *International conference on machine learning*, 4615–4625. PMLR.

Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 1–7.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.

Wang, G.; Payani, A.; Lee, M.; and Kompella, R. 2023. Mitigating group bias in federated learning: Beyond local fairness. *arXiv preprint arXiv:2305.09931*.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and Optimal Classification via Post-Processing. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 37977–38012. PMLR.

Yuan, H.; Zaheer, M.; and Reddi, S. 2021. Federated composite optimization. In *International Conference on Machine Learning*, 12253–12266. PMLR.

Zeng, Y.; Chen, H.; and Lee, K. 2021. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*.

Zhang, Y.; Yin, Z.; Li, Y.; Yin, G.; Yan, J.; Shao, J.; and Liu, Z. 2020. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 70–85. Springer.