Multi-Dimensional Fair Federated Learning

Cong Su[1], Guoxian Yu[1,2,*], Jun Wang[2], Hui Li[1,2], Qingzhong Li[1,2,*], Han Yu[3]

[1]School of Software, Shandong University, Jinan, China
[2]SDU-NTU Joint Centre for AI Research, Shandong University, Jinan, China
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
csu@mail.sdu.edu.cn, {gxyu, kingjun, lih, lqz}@sdu.edu.cn, han.yu@ntu.edu.sg

Abstract

Federated learning (FL) has emerged as a promising collaborative and secure paradigm for training a model from decentralized data without compromising privacy. *Group fairness* and *client fairness* are two dimensions of fairness that are important for FL. Standard FL can result in disproportionate disadvantages for certain clients, and it still faces the challenge of treating different groups equitably in a population. The problem of privately training fair FL models without compromising the generalization capability of disadvantaged clients remains open. In this paper, we propose a method, called `mFairFL`, to address this problem and achieve group fairness and client fairness simultaneously. `mFairFL` leverages differential multipliers to construct an optimization objective for empirical risk minimization with fairness constraints. Before aggregating locally trained models, it first detects conflicts among their gradients, and then iteratively curates the direction and magnitude of gradients to mitigate these conflicts. Theoretical analysis proves `mFairFL` facilitates the fairness in model development. The experimental evaluations based on three benchmark datasets show significant advantages of `mFairFL` compared to seven state-of-the-art baselines.

Introduction

The widespread adoption of machine learning models has given rise to significant apprehensions regarding fairness, spurring the emergence of fairness criteria and models. In recent times, a multitude of fairness criteria have been put forth, with one of the most widely acknowledged ones being **group fairness** (Hardt, Price, and Srebro 2016; Ustun, Liu, and Parkes 2019). Group fairness might also be mandated by legal statutes (EU et al. 2012), necessitating models to impartially treat distinct groups concerning sensitive attributes such as age, gender, and race. Building upon these concepts of group fairness, numerous methodologies have been introduced to train equitable models, predicated on the premise that the model can directly access the complete training dataset (Zafar et al. 2017; Roh et al. 2021). However, the ownership of these datasets often resides with disparate institutions, rendering them inaccessible for sharing due to privacy safeguarding considerations.

*Corresponding author: Guoxian Yu and Qingzhong Li.

Federated learning (FL) (Wang et al. 2021a) stands as a distributed learning paradigm that facilitates the collective training of a model by multiple data owners, all while retaining their data within their local domains. If each data owner was to individually train a fairness model on their own data and subsequently contribute it for aggregation, akin to the methods of FedAvg (McMahan et al. 2017) and FedOPT (Reddi et al. 2020), a promising way opens up for augmenting model fairness within decentralized contexts. However, the presence of data heterogeneity, manifesting in variations in sizes and distributions across different clients, introduces a distortion to the localized efforts aimed at enhancing fairness in the global model.

Consequently, a disparity emerges between the impartial model aggregated in a straightforward manner, utilizing fairness models trained on diverse client datasets, and the model achieved under centralized circumstances. Meanwhile, a simplistic pursuit of minimizing the aggregation loss in the federated system can lead the global model astray, favoring certain clients excessively and disadvantaging others, thereby engendering what is termed as *client fairness* (CF). Preceding endeavors have predominantly centered on rectifying issues concerning client fairness. These efforts encompass methodologies such as re-weighting client aggregation weights (Zhao and Joshi 2022), tackling distributed mini-max optimization challenges (Mohri, Sivek, and Suresh 2019), or mitigating conflicts between clients (Hu et al. 2022).

In contrast, our emphasis pivots toward multi-dimensional fairness, encompassing both group and client fairness, aligning with legal stipulations and ethical considerations. This dual focus also significantly influences the willingness of clients to actively engage in the FL process, thereby contributing to datasets that are more comprehensive and representative for the training of the global model. However, the inherent decentralized nature of this approach provokes complexities in achieving equitable training for a global model, particularly when confronted with the complexities of heterogeneous data distributions spanning the client landscape. The intricate challenge of privately training an equitable model from such decentralized, disparate data, while ensuring equitable treatment for each contributing client, poses a formidable conundrum. We aim to address this open and intricate quandary.

We propose the Multi-dimensional Fair Federated Learning (mFairFL) method, which aims to ensure equity not only among distinct sensitive groups but also across individual clients. The core principle of mFairFL involves the optimization of client models under the guidance of fairness constraints. Prior to the execution of gradient aggregation on the central server, mFairFL evaluates the potential presence of conflicting gradients among clients by assessing their gradient similarities. Subsequently, mFairFL undertakes an iterative process wherein it tactically adjusts the direction and magnitude of conflicting gradients to mitigate such disparities. Through this nuanced strategy, mFairFL adeptly navigates the delicate balance between equitable treatment and optimal accuracy, catering to both marginalized sensitive groups and individual clients. The schematic framework of mFairFL is depicted in Figure 1. Our contributions can be succinctly outlined as follows:

(i) We introduce an innovative framework for fair federated learning, denoted as mFairFL, and establish its capacity to bolster model fairness concerning sensitive groups within a decentralized data context.

(ii) mFairFL conceptualizes the pursuit of fairness optimization through a meticulously designed minimax framework, replete with a group fairness metric as constraints. It analyzes and adjusts the trajectory and magnitude of potentially conflicting gradients throughout the training process, which adeptly augments group fairness across the entire populace while ensuring an impartial treatment of each client within the global model.

(iii) Through both theoretical and experimental analysis, we demonstrate that mFairFL excels in mitigating gradient conflicts among clients, ultimately achieving a higher degree of group fairness compared to the state of the art (SOTA).

Related Work

With the growing concern surrounding fairness, various approaches have been proposed. To analyze the problem, we categorize fairness models into two types: centralized and federated, based on their training protocols.

Fairness models on centralized data. In the context of centralized data, it is common to modify the training framework to attain an appropriate level of group fairness, ensuring that a classifier exhibits comparable performance across different sensitive groups. Several techniques have been devised to address group fairness issues within the centralized setting, which can be categorized into three types: pre-processing (Salimi et al. 2019), in-processing (Garg et al. 2019), and post-processing (Mishler, Kennedy, and Chouldechova 2021) methods. For more extensive insights into fairness methods applied to centralized data, refer to the recent literature survey (Pessach and Shmueli 2022).

Fairness FL models on decentralized data. In contrast, achieving fairness within the practical FL setting has received limited attention compared to centralized solutions (Wang et al. 2021b). The notion of 'fairness' in FL differs slightly from the standard concept in centralized learning. *Client fairness*, a popular fairness definition in FL, aims to ensure that all clients (i.e., data owners) achieve similar accuracy. Previous attempts to achieve client fairness in FL include modifying the aggregation weights of clients to achieve a uniform service quality for all clients (Li et al. 2019; Zhao and Joshi 2022; Yue, Nouiehed, and Al Kontar 2023), selecting biased clients for update (Cho et al. 2020), and finding the common update direction for all clients (Hu et al. 2022). Only a few studies are dedicated to group fairness in FL (Abay et al. 2020; Du et al. 2021; Gálvez et al. 2021). For example, Zeng, Chen, and Lee (2021) updated the weight of local loss for each sensitive group during the global aggregation phase. Ezzeldin et al. (2023) adapted client weights based on local fairness of each client and deviations from global one. However, these methods disregard the gradient conflicts, which lead to performance decline and unfavourable outcomes for certain clients.

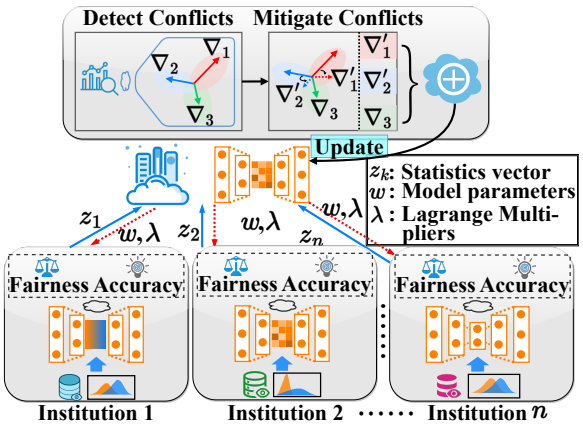

Figure 1: An overview of mFairFL, which formulates a minimax constrained optimization problem in terms of accuracy and fairness. Before aggregating client gradients, it detects the presence of gradient conflicts, and then mitigates the conflicting gradients through gradient adjustments to align them with the overall fairness objective.

mFairFL aims to eliminate bias towards different groups (group fairness) based on sensitive attributes and to learn a global model that benefits all clients, thereby achieving better client fairness alongside group fairness.

Preliminaries

Federated Learning

Following the typical FL setting (McMahan et al. 2017), suppose there are K different clients, and each client can only access its own dataset $\mathcal{D}_k = \{d_k^i = (s_k^i, x_k^i, y_k^i)\}_{i=1}^{n_k} \in \mathcal{D}$, where s_k^i is the sensitive attribute of client k, y_k^i is the label, x_k^i is other observational attributes, n_k is the number of client samples. The goal of FL is to train a global model parameterized with $w \in \mathbb{R}^m$ (m is the number of parameters) on client datasets $\mathcal{D}_k$ with guaranteed privacy:

$$\min_{w \in \mathbb{R}^m} \sum_{i=1}^{K} p_i L(\mathcal{D}_i, w) \tag{1}$$

where $L(\mathcal{D}_i, w) = \frac{1}{n_i} \sum_{i=1}^{n_i} l(d_k^i, w)$ is the local objective function of client i with weights $p_i \geq 0$, $\sum_{i=1}^{K} p_i = 1$.

Fairness Notions

In this paper, We focus on three canonical group fairness notions, i.e., *Demographic Parity* (DP), *Equalized Odds* (EO), and *Accuracy Parity* (AP) (Pessach and Shmueli 2022). For the sake of exposition, we describe these notions in the centralized setting.

Definition 1 (Demographic Parity) *The model's predictions $\hat{Y}$=$\hat{y}$ are statistically independent of the sensitive attribute S. The extent of a model's unfairness with respect to* Demographic Parity *can be measured as follows:*

$$DP(\hat{y}) = |\mathbb{P}[\hat{Y} = \hat{y}|S = s] - \mathbb{P}[\hat{Y} = \hat{y}]| \; \forall s \in S \tag{2}$$

Definition 2 (Equalized Odds) *Given the label Y=y, the predictions $\hat{Y}$=$\hat{y}$ are statistically independent of the sensitive attribute S. i.e., for all $s \in S$ and $y \in Y$, we can measure the absolute difference between two prediction rates to quantify how unfair a model is in term of* Equalized Odds*:*

$$EO(\hat{y}) = |\mathbb{P}[\hat{y}|S = s, Y = y] - \mathbb{P}[\hat{y}|Y = y]| \tag{3}$$

Definition 3 (Accuracy Parity) *The model's misprediction rate is independent of the sensitive attribute:*

$$\mathbb{P}[\hat{y} \neq y|S = s] = \mathbb{P}[\hat{y} \neq y] \; \forall s \in S \tag{4}$$

where, equivalently, we can measure the degree of unfairness in the model with respect to Accuracy Parity *as follows:*

$$AP(\hat{y}) = |\mathbb{P}[L(\mathcal{D}, w)|S = s] - \mathbb{P}[L(\mathcal{D}, w)]| \tag{5}$$

where $L(\mathcal{D}, w)$ is the loss function minimized in problem (1)*.*

The above discussed fairness notions can be interpreted as the difference between each group and the overall population (Fioretto, Mak, and Van Hentenryck 2020). Formally, these notions can be rewritten as:

$$FN = |F(\mathcal{D}, w) - F(\mathcal{D}^s, w)| \tag{6}$$

where $F(\mathcal{D}, w) = \frac{1}{n}\sum_{d\in\mathcal{D}} f(d, w)$, $F(\mathcal{D}^s, w) = \frac{1}{n_s}\sum_{d^s\in\mathcal{D}} f(d^s, w)$. $\mathcal{D}^s$ is the subset of $\mathcal{D}$ with S=s, and f is one of the fairness notions described above.

Necessity for FL to Improve Fairness

In this subsection, we analyse the advantage of FL for improving fairness in decentralized settings. To build a fair model in decentralized settings, an intuitive solution (hereon referred to as "IndFair") is to independently train the fair local model using client data. Specifically, for client k, IndFair trains a fair model by solving the following problem:

$$\begin{aligned} &\min L(\mathcal{D}_k, w) \\ &s.t. \; |F(\mathcal{D}_k, w) - F(\mathcal{D}_k^s, w)| \leq \alpha, \forall s \in S \end{aligned} \tag{7}$$

where α∈$[0, 1]$ is the fairness tolerance threshold. Let g_k^α be the trained model of client k, then the overall performance of IndFair is defined as the mixture of all clients:

$$\begin{aligned} \hat{\mathrm{y}}|\mathrm{x}, \mathrm{s} &\sim \begin{cases} \mathrm{Bern}(g_1^\alpha(\mathrm{x}, \mathrm{s})), & \text{w.p. } 1/K \\ \mathrm{Bern}(g_2^\alpha(\mathrm{x}, \mathrm{s})), & \text{w.p. } 1/K \\ \cdots\cdots \\ \mathrm{Bern}(g_K^\alpha(\mathrm{x}, \mathrm{s})), & \text{w.p. } 1/K \end{cases} \\ &= \mathrm{Bern}((g_1^\alpha(\mathrm{x}, \mathrm{s}) + \cdots + g_K^\alpha(\mathrm{x}, \mathrm{s}))/K) \\ &= \mathrm{Bern}(g_\alpha^{Seq}) \end{aligned} \tag{8}$$

where Bern stands for Bernoulli distribution, and $\mathrm{w.p.}$ is the abbreviation for 'with probability'.

On the other hand, we can train a fair global model (hereon referred to as "FedFair") on decentralized data through FL. The fair global model g_α^{Fed} is obtained by solving a constrained problem:

$$\begin{aligned} &\min L(\mathcal{D}, w), \\ &s.t. \; |F(\mathcal{D}_k, w) - F(\mathcal{D}_k^s, w)| \leq \alpha, \\ &\quad \textit{for all} \quad k = 1, 2, ..., K. \end{aligned} \tag{9}$$

Here, an important question raises: *can FedFair achieve a better fairness than IndFair?* The following theorem gives the confirm answer.

Theorem 1 (Necessity for FL) *If the data distribution is highly heterogeneous across clients, then* $\min FN(g_\alpha^{Ind}) > \min FN(g_\alpha^{Fed})$.

Theorem 1 means that in decentralized setting, there is a fairness gap between federated methods and non-federated ones, and FL improves the fairness performance. The proof and formal representation are deferred into the Supplementary file (Su et al. 2023).

The Proposed Approach

Theorem 1 demonstrates the potency of FL in effectively bolstering model fairness while safeguarding against data leakage within a decentralized context. Nevertheless, employing fairness methods directly within the FL framework might not be the optimal approach. This challenge arises from the significant heterogeneity in data distributions across clients. Consequently, the localized fairness performance could diverge from fairness across the entire population. Additionally, in this scenario, the concept of client fairness gains prominence as another critical facet of fairness that necessitates consideration.

To tackle these intricacies, we introduce `mFairFL`, a solution designed to confidentially train a global model while integrating group fairness. This approach effectively mitigates the adverse effects of gradient conflicts among clients, as depicted in Figure 1. `mFairFL` strategically transforms the fairness-constrained problem into an unconstrained problem that enforces fairness through the use of Lagrange multipliers. In each communication round, every client computes its training loss, measures of fairness violations, and gradients. Subsequently, these statistics are communicated to the FL server (aggregation phase). The server then identifies and rectifies conflicting gradients' direction and magnitude before aggregation. This refined model is then updated and distributed to clients (local training phase).

This intricate process enables mFairFL to attain a precise global model that remains equitable for both sensitive groups and individual clients. The subsequent subsections delve into the finer technical intricacies of our approach.

Local Statistics Computation Phase

Our goal is to train an optimal model from decentralized data while satisfying group fairness. For this purpose, we directly inject the group fairness constraint into the model training:

$$\begin{aligned} &w^* = \min_{w \in \mathbb{R}^m} L(\mathcal{D}, w) \\ &s.t.\ |F(\mathcal{D}, w) - F(\mathcal{D}^s, w)| \leq \alpha, \forall s \in S \end{aligned} \tag{10}$$

where $L(\mathcal{D}, w) = \frac{1}{n} \sum_{d \in \mathcal{D}} l(d, w)$, $F(\mathcal{D}, w)$ is the fairness metric defined in Eq. (6).

Let $\mathbf{h}(w)$=$[h_1(w), h_2(w), \cdots, h_{|S|}(w)]$ where $h_s(w)$=$|F(\mathcal{D}, w) - F(\mathcal{D}^s, w)| - \alpha$. We use the similar technique from the Lagrangian approach (Fioretto et al. 2021) to relax the constraint:

$$J(w, \alpha) = L(\mathcal{D}, w) + \lambda \mathbf{h}(w). \tag{11}$$

The relaxation provides more freedom for the optimization algorithm to find solutions that may not strictly satisfy all the constraints, but rather approximate them within an acceptable range.

Thus, the objective function in Eq. (11) can be optimized using gradient descent/ascent:

$$\begin{cases} \lambda \leftarrow \lambda + \gamma \mathbf{h}(w), \\ w \leftarrow w - \eta(\nabla_w L(\mathcal{D}, w) + \lambda \nabla_w \mathbf{h}(w)). \end{cases} \tag{12}$$

Based on Eq. (12), each client computes the following statistics required for the server to perform model updates:

$$\begin{aligned} &L(\mathcal{D}, w); \nabla_w L(\mathcal{D}, w); F(\mathcal{D}, w); [F(\mathcal{D}^s, w)]_{s \in S}; \\ &\nabla_w F(\mathcal{D}, w); \nabla_w [F(\mathcal{D}^s, w)]_{s \in S}. \end{aligned} \tag{13}$$

In fact, some of these statistics can be obtained from others: $F(\mathcal{D}, w)$=$\sum_{s \in S} F(\mathcal{D}^s, w)$, $\nabla_w F(\mathcal{D}, w)$=$\sum_{s \in S} \nabla_w F(\mathcal{D}^s, w)$. Therefore, in each communication round, a client reports a statistics vector to the server as:

$$\begin{aligned} \mathbf{z}_k = &[L(\mathcal{D}_k, w); \nabla_w L(\mathcal{D}_k, w); [F(\mathcal{D}_k^s, w)]_{s \in S}; \\ &\nabla_w [F(\mathcal{D}_k^s, w)]_{s \in S}] \end{aligned} \tag{14}$$

We define the training loss of client k in round t as l_k^t=$L(\mathcal{D}_k, w) + \lambda \mathbf{h}(w)$, and the updated gradient g_k^t=$\nabla_w L(\mathcal{D}_k, w) + \lambda \nabla_w \mathbf{h}(w)$. Let G_t=$\{g_1^t, g_2^t, \cdots, g_K^t\}$ represent the gradients received by the server from clients, and L_t=$\{l_1^t, l_2^t, \cdots, l_K^t\}$ be the received client losses.

Aggregation Phase

During the aggregation phase, the server leverages $\mathbf{z}_k$ provided by clients to refine and update the global model via Eq. (12). Owing to the presence of diverse data distributions, gradient conflicts emerge among clients. In isolation, these conflicts might not be inherently detrimental, as straightforward gradient averaging can effectively optimize the global objective function (McMahan et al. 2017). However, when conflicts among gradients involve considerable variations in magnitudes, certain clients could encounter pronounced drops in performance. For instance, consider the scenario of training a binary classifier. If a subset of clients holds a majority of data pertaining to one class, and conflicts in gradients arise between these two classes, the global model could become skewed toward the majority-class clients, thereby compromising performance on the other class. Moreover, even when class balance is maintained among clients, disparities in gradient magnitudes may persist due to divergent sample sizes across clients.

Therefore, before aggregating clients' gradients in each communication round, mFairFL first checks whether there are any conflicting gradients among clients. If there are gradient conflicts, then at least a pair of client gradients (g_i^t, g_j^t) such that $\cos(g_i^t, g_j^t) < \hat{\phi}_{ij}^t$, where $\hat{\phi}_{ij}^t \geq 0$ is the gradient similarity goal of t-th communication round. Note that interactions among gradients (gradient similarity goal) change significantly across clients and communication rounds. Thus, mFairFL performs Exponential Moving Average (EMA) (Wang and Tsvetkov 2021) to set appropriate gradient similarity goals for clients i and j in round t:

$$\hat{\phi}_{ij}^t = \delta \hat{\phi}_{ij}^{t-1} + (1 - \delta) \phi_{ij}^t \tag{15}$$

where δ is the hyper-parameter, and $\phi_{ij}^t = \cos(g_i^t, g_j^t)$ is the computed gradient similarity. Specifically, $\hat{\phi}_{ij}^0 = 0$.

In order to mitigate the adverse repercussions stemming from gradient conflicts among clients, mFairFL introduces an innovative gradient aggregation strategy. Specifically, the approach initiates by arranging clients' gradients within G_t in ascending order, based on their respective loss values. This orchestrated arrangement yields PO_t, which outlines the sequence for utilizing each gradient as a reference projection target. Subsequently, through an iterative process, mFairFL systematically adjusts the magnitude and orientation of the k-th client gradient, denoted as g_k^t, so as to align with the desired similarity criteria between g_k^t and the target gradient $g_j^t \in PO_t$, in accordance with the prescribed order set by PO_t:

$$g_k^t = c_1 \cdot g_k^t + c_2 \cdot g_j^t \tag{16}$$

Since there are infinite valid combinations of c_1 and c_2, we fix $c_1 = 1$ and apply the Law of Sines on the planes of g_k^t and g_j^t to calculate the value of c_2, and obtain the derived new gradient for the k-th client:

$$g_k^t = g_k^t - \frac{||g_k^t||(\phi_{kj}^t \sqrt{1 - (\hat{\phi}_{kj}^t)^2} - \hat{\phi}_{kj}^t \sqrt{1 - (\phi_{kj}^t)^2})}{||g_j^t|| \sqrt{1 - (\hat{\phi}_{kj}^t)^2}} \cdot g_j^t \tag{17}$$

The derivation detail is deferred into the Supplementary file.

Theorem 2 *Suppose there is a set of gradients* $G = \{g_1, g_2, ..., g_K\}$ *where* g_i *always conflicts with* $g_j^{t_j}$ *before adjusting* $g_j^{t_j}$ *to match similarity goal between* $g_j^{t_j}$ *and* g_i *(*$g_j^{t_j}$ *represents the gradient adjusting* g_j *with the target gradients in G for* t_j *times). Suppose* $\epsilon_1 \leq |cos(g_i^{t_i}, g_j^{t_j})| \leq \epsilon_2$,

$0<\epsilon_1<\hat{\phi}_{ij}\leq\epsilon_2\leq 1$, *for each* $g_i \in G$, *as long as we iteratively project* g_i *onto* g_k*'s normal plane (skipping* g^i *itself) in the ascending order of* k=$1, 2, \cdots, K$, *the larger the* k *is, the smaller the upper bound of conflicts between the aggregation gradient of global model* g^{global} *and* g_k *is. The maximum value of* $|g^{global} \cdot g_k|$ *is bounded by* $\frac{K-1}{K}(\max_i ||g_i||)^2 \frac{\epsilon_2 X_{\max}(1-X_{\min})(1-(1-X_{min})^{K-k})}{X_{\min}}$, *where* $X_{\max}=\frac{\epsilon_2\sqrt{1-\hat{\phi}^2}-\hat{\phi}\sqrt{1-\epsilon_2^2}}{\sqrt{1-\hat{\phi}^2}}$ *and* $X_{\min}=\frac{\epsilon_1\sqrt{1-\hat{\phi}^2}-\hat{\phi}\sqrt{1-\epsilon_1^2}}{\sqrt{1-\hat{\phi}^2}}$.

Theorem 2 substantiates that the later a client's gradient assumes the role of the projection target, the fewer conflicts it will engage in with the ultimate averaged gradient computed by `mFairFL`. Consequently, in the pursuit of refining the model's performance across clients with comparatively lower training proficiency, we position clients with higher training losses towards the end of the projecting target order list, denoted as PO_t. Additionally, these gradients provide the optimal model update direction. To further amplify the focus on *client fairness*, we permit βK clients with suboptimal performance to retain their original gradients. The parameter β modulates the extent of conflict mitigation and offers a means to strike a balance. When β=1, all clients are mandated to mitigate conflicts with others. Conversely, when β=0, all clients preserve their original gradients, aligning `mFairFL` with FedAvg. By adopting this approach, `mFairFL` effectively alleviates gradient conflicts, corroborated by the findings in Theorem 2. Consequently, `mFairFL` is equipped to set an upper limit on the maximum conflict between any client's gradient and the aggregated gradient of the global model. This strategic stance enables `mFairFL` to systematically counteract the detrimental repercussions stemming from gradient conflicts. Algorithm 1 in the Supplementary file outlines the main procedures of `mFairFL`.

Theorem 3 proves that `mFairFL` can find the optimal value w^* within a finite number of communications. This explains why `mFairFL` can effectively train a group and client fairness-aware model in the decentralized setting. The proof can be found into the Supplementary file.

Theorem 3 *Suppose there are* K *objective functions* $J_1(w), J_2(w), \cdots, J_K(w)$*, and each objective function is differentiable and L-smooth. Then* `mFairFL` *will converage to the optimal* w^* *within a finite number of steps.*

Experimental Evaluation

Experimental Setup

In this section, we conduct experiments to evaluate the effectiveness of `mFairFL` using three real-world datasets: Adult (Dua and Graff 2017), COMPAS (ProPublica. 2016), and Bank (Moro, Cortez, and Rita 2014). The Adult dataset contains 48,842 samples, with 'gender' treated as the sensitive attribute. There are 7,214 samples in the COMPAS dataset, with 'gender' treated as the sensitive attribute. As for the Bank dataset with 45,211 samples, with 'age' treated as the sensitive attribute. We split the data among five FL clients in an non-iid manner.[1]

For the purpose of comparative analysis, we consider several baseline methods, categorized into three groups: (i) independent training of the fair model within a decentralized context (IndFair); (ii) fair model training via FedAvg (FedAvg-f); (iii) fair model training within a centralized setting (CenFair). Three SOTA FL with group fairness: (i) FedFB (Zeng, Chen, and Lee 2021), which adjusts each sensitive group's weight for aggregation; (ii) FPFL (Gálvez et al. 2021), which enforces fairness by solving the constrained optimization; (iii) FairFed (Ezzeldin et al. 2023), which adjusts clients' weights based on locally and global trends of fairness mtrics. In addition to these, we evaluate our proposed `mFairFL` against cutting-edge FL methods that emphasize *client fairness*, including: (i) q-FFL (Li et al. 2019), which adjusts client aggregation weights using a hyperparameter q; (ii) DRFL (Zhao and Joshi 2022), which automatically adapts client weights during model aggregation; (iii) Ditto (Li et al. 2021), a hybrid approach that merges multitask learning with FL to develop personalized models for each client; and (iv) FedMGDA+ (Hu et al. 2022), which frames FL as a multi-objective optimization problem. Throughout our experiments, we adhere to a uniform protocol of 10 communication rounds and 20 local epochs for all FL algorithms. For other methods, we execute 200 epochs, leveraging cross-validation techniques on the training sets to determine optimal hyperparameters for the comparative methods. All algorithms are grounded in ReLU neural networks with four hidden layers, thereby ensuring an equal count of model parameters.[2] We use the same server (Ubuntu 18.04.5, Intel Xeon Gold 6248R and Nvidia RTX 3090) to perform experiments.

Estimation on Group Fairness

We undertake a comprehensive comparative analysis, focusing on the accuracy and group fairness aspects of the evaluated methods. To delve into the intricate relationship between method performance and data heterogeneity, we extend the setting of McMahan et al. (2017) for constructing heterogeneous data, emphasizing the heterogeneity in the sensitive attributes and data sizes across clients. Specifically, we group the datsets by sensitive attributes, and randomly assign 30%, 30%, 20%, 10%, 10% of the samples from group 0 and 10%, 20%, 20%, 20%, 30% of the samples from group 1 to five clients, respectively. The outcomes of this data splitting strategy, encompassing average accuracy along with standard deviations and the Demographic Parity violation score for each method, are outlined in Table S2. Furthermore, for datasets characterized by pronounced data heterogeneity, we draw samples from each group across five clients at a ratio of 50%, 10%, 10%, 20%, 10%, and 10%, 40%, 30%, 10%, 10%, respectively. The corresponding experimental outcomes are showcased in Table 1. From the

[1]Due to page limit, we include the experiments conducted in a more general setting with multiple sensitive attributes and multiple values for sensitive attributes in Supplementary file (In Table S3).

[2]Further elaboration on the selection of hyperparameters for `mFairFL` can be found in the Supplementary file.

	Adult					Compas					Bank				
	Acc.	DP	EO	AP	CF	Acc.	DP	EO	AP	CF	Acc.	DP	EO	AP	CF
IndFair	.768●	.083●	.071●	.077●	-	.573●	.083●	.097●	.085●	-	.831	.028●	.025●	.029●	-
FedAvg-F	.706●	.224●	.164●	.218●	.232●	.558●	.059●	.066●	.062●	.184●	.828●	.033●	.034●	.033●	.143●
FedFB	.779	.014	.007	.011	.058●	.557●	.023●	.019●	.021●	.033	.837	.014●	.016●	.009	.067●
FPFL	.754●	.023●	.016●	.019●	.228●	.553●	.033●	.018●	.024●	.157●	.822●	.008	.012●	.010●	.153●
FairFed	.756●	**.009**	.004	.008	.244●	.551●	.009	**.003**	.004	.186●	.824●	.003	.004	.004	.164●
FedMGDA+	**.837**○	.238●	.237●	.238●	.063●	**.635**○	.136●	.141●	.137●	.044●	**.874**○	.084●	.077●	.085●	.065●
CenFL	.812○	.014	.008	.013	-	.616	.014●	.008	.011●	-	.866○	**.001**	**.000**	**.002**	-
mFairFL	.792	.012	**.003**	**.007**	**.036**	.596	**.005**	.009	**.003**	**.022**	.844	.005	.006	.003	**.028**

Table 1: Accuracy and the violation of Group fairness and Client Fairness results on the three datasets with high data heterogeneity among clients. The best results in fairness are highlighted in boldface. ○/● indicates that mFairFL is statistically worse/better than the compared method by student pairwise t-test at 95% confident level. '-' implies not applicable.

insights gleaned from Tables S2 and 1, we observe that:
(i) mFairFL prominently enhances fairness, achieving a parity of fairness akin to CenFair. This substantiates mFairFL's efficacy in skillfully training fair models for sensitive groups within the decentralized data landscape.
(ii) The lackluster performance of IndFair in terms of group fairness accentuates that in a decentralized scenario, fairness models exclusively trained on local data fall considerably short of achieving group fairness at a population-wide level. It is also noteworthy that FedAvg-f occasionally exhibits lower accuracy than IndFair. This discrepancy arises from the aggregation strategy of FedAvg-f, which can have unintended consequences for certain clients. Conversely, IndFair manages to ensure fairness for specific sub-distributions through the training of each local model.

(iii) In direct comparison, mFairFL distinctly outperforms FedAvg-f in both accuracy and fairness, thus underscoring the constraints inherent in merely grafting fairness techniques onto the FL paradigm. Evidently, the group fairness achieved by FedAvg-f lags behind the fairness exhibited across the entire population. This gap is particularly pronounced in scenarios characterized by high data heterogeneity among clients. Through the judicious amalgamation of fairness techniques with the decentralized essence of FL, and its steadfast commitment to ensuring advantageous model updates for all clients, mFairFL adeptly enhances both the overarching fairness and accuracy, thereby offering a comprehensive improvement.

(iv) Notably, mFairFL can better trade-off accuracy and group fairness than FedFB, FPFL and FairFed. This is because they overlook the detrimental effects of the conflicting gradients with large difference in the magnitudes, leading to accuracy reduction and harming certain clients. FedMGDA+ frequently yields the highest accuracy in various scenarios, but markedly infringes upon the fairness of model decisions as applied to disadvantaged groups. This is primarily attributed to the fact that FedMGDA+ concentrate solely on aligning client accuracy without due regard for mitigating discrimination against sensitive groups.

(v) Upon juxtaposing the outcomes presented in Table 1 (high heterogeneity) with those in Table S2 (low heterogeneity), a salient observation arises: mFairFL demonstrates a marginal decrease in both fairness and accuracy when transitioning from low to high data heterogeneity. This underscores the robustness intrinsic to mFairFL when grappling with heterogeneous data. Such a consistency aligns with our initial expectations, as mFairFL adeptly orchestrates gradient directions and magnitudes to navigate conflicts, thereby ensuring equitable model updates across all clients. Conversely, FedAvg-f manifests notable performance disparities across distinct data heterogeneity levels, with a particularly steep decline observed in scenarios characterized by high data heterogeneity. This vulnerability is attributed to FedAvg-f's simplistic gradient averaging approach, which insufficiently accommodates the intricate impact of data heterogeneity on the global model.

Estimation on Client Fairness

The group fairness-aware model cultivated by mFairFL brings about advantages for each participating client, all while avoiding any undue preference towards specific clients. To further validate this assertion, we embark on a series of experiments designed to evaluate mFairFL's performance in terms of *Client Fairness*, subsequently juxtaposing it against other pertinent fairness methods. *Client Fairness* stands as a potent metric for gauging whether the global model disproportionately favors particular clients while disregarding the rest. To further accentuate the discerning capabilities of mFairFL, we undertake the random allocation of samples: 50% and 10% of group 0 samples, coupled with 10%, 20%, and 10% of group 1 samples, are assigned to the 1st, 2nd, 3rd, 4th, and 5th clients, respectively. This deliberate strategy accentuates pronounced data heterogeneity across clients. For reference, FedAvg constitutes the baseline in this experimental setup. The resulting accuracy and violation scores pertinent to *Client Fairness* for each method are succinctly presented in Table 2. Our observations from this comparative analysis are as follows:
(i) Remarkably, mFairFL emerges as the frontrunner, boasting the most modest client fairness violation scores while achieving accuracy on par with other fairness-focused FL methods. In essence, mFairFL excels in abating the potential biases inherent to FL contexts. Through its meticulous consideration of conflicting gradients and adept adjustments to their directions and magnitudes, mFairFL guarantees a more equitable distribution of model updates among

		FedAvg	q-FFL	DRFL	Ditto	FedMGDA+	mFairFL
Adult	Accuracy	.792•	.718•	.762•	.834•	**.837**	.853
	CF Vio.	.219•	.081•	.084•	.074•	.063•	**.035**
Compas	Accuracy	.594•	.566•	.598•	.629•	.635•	**.668**
	CF Vio.	.179•	.058•	.031•	.024	.044•	**.018**
Bank	Accuracy	.842•	.816•	.864•	.877	.874•	**.889**
	CF Vio.	.147•	.110•	.090•	.068•	.065•	**.022**

Table 2: Accuracy (↑) and Client Fairness violation score (↓) on three datasets with high heterogeneity among clients. ∘/• indicates that mFairFL is statistical worse/better than the compared method (student pairwise t-test at 95% confident level).

clients. This concerted effort tangibly diminishes the breach of client fairness, ultimately heralding a more even allocation of model updates among all participating clients. In contrast, both q-FFL and DRFL endeavor to tackle client fairness by manipulating client aggregation weights, but falter in effectively addressing conflicts characterized by substantial gradient magnitude disparities. Ditto aims to strike a balance between local and global models, engendering personalized models for individual clients. However, its global model aggregation strategy closely resembles that of FedAvg, potentially yielding unfavorable outcomes for certain clients. In the same vein, FedMDGA+ aspires to pinpoint a shared update direction for all clients during federated training, inadvertently overlooking the influential role played by gradient magnitudes in model aggregation. Therefore, it is evident that mFairFL stands as the epitome of achievement, outperforming its counterparts both in terms of client fairness and accuracy.

(ii) FedAvg, unfortunately, languishes at the bottom of the performance spectrum, marked by inferior accuracy and client fairness. This regression is traceable to its rudimentary averaging strategy, which disregards the disparate contributions of individual clients. Consequently, when confronting gradients in conflict with significantly divergent magnitudes, FedAvg becomes susceptible to overfitting certain clients at the detriment of others. The significant difference in performance between mFairFL and FedAvg demonstrates the potency of mFairFL in counteracting client conflicts.

(iii) To further solidify mFairFL's efficacy, we furnish the number of iterations imperative for all compared methods to attain their optimal performance in Figure S1 of the Supplementary file. This visual depiction affords insights into the convergence trajectories undertaken by distinct methods over iterations. Notably, mFairFL exhibits commendable performance levels and converges towards the pinnacle of client fairness within a noticeably fewer (or comparable) count of communication rounds. This clearly underscores mFairFL's capacity to efficiently train the model, attaining the desired accuracy and client fairness benchmarks with commendable efficacy.

Ablation Study

To prove the necessity of the projection order of mFairFL, we introduce two variants of mFairFL: (i) mFairFL-rnd adjusts the gradients in a random order of the projection target. (ii) mFairFL-rev adjusts gradients in the opposite order. The experimental settings are the same as the previous subsection. The results are shown in Table 3. It can be observed that the projection order has significant impact on the effectiveness of gradient projection. mFairFL-rnd ignores the information provided by client losses. Thus, it loses to mFairFL in terms of *group fairness* and *client fairness*. mFairFL-rev achieves lower fairness than mFairFL-rnd, indicating that the global model tends to neglecting clients with poorer performance when adjusting gradients in the opposite order of mFairFL. The best multi-dimensional fairness performance is obtained by mFairFL. This confirms that its loss-based order helps improve fairness.

		ours-rnd	ours-rev	ours
Adult	Acc.	.774	.768•	**.792**
	DP.	.018	.027•	**.012**
	EO	.013•	.026•	**.003**
	AP	.017•	.028•	**.007**
	CF	.049•	.064•	**.036**
Compas	Acc.	.577•	.580	**.596**
	DP	.014•	.020•	**.005**
	EO	.015•	.018•	**.009**
	AP	.012•	.019•	**.003**
	CF	.044•	.049•	**.022**
Bank	Acc.	.837	.822•	**.844**
	DP	.009	.023•	**.005**
	EO	.013	.019•	**.006**
	AP	.007	.021•	**.003**
	CF	.047•	.077•	**.028**

Table 3: Accuracy (↑), Group fairness and Client Fairness violation scores (↓) of mFairFL and its variants. • indicates that mFairFL is statistical better than the variant.

Conclusions

Addressing both group fairness and client fairness is paramount in the realm of FL. This paper introduces the novel mFairFL method as a groundbreaking solution that adeptly navigates these dual dimensions of fairness. mFairFL formulates the optimization conundrum as a minimax problem featuring group fairness constraints. Through meticulous adjustments to conflicting gradients throughout the training regimen, mFairFL orchestrates model updates that distinctly benefit all clients in an equitable manner. Both theoretical study and empirical results confirm that mitigating client conflicts during global model update improves the fairness for sensitive groups, and mFairFL effectively achieves both group fairness and client fairness. How to mitigate the privacy risks of mFairFL is our future work.

Acknowledgements

This work is supported by NSFC (62031003, 62072380 and 62272276); National Key Research and Development Program of China (2022YFC3502101); Taishan Scholars Project Special Funding; Xiaomi Young Talents Program; CAAI-Huawei MindSpore Open Fund; the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019); and the RIE 2020 Advanced Manufacturing and Engineering (AME) Programmatic Fund (No. A20G8b0102), Singapore.

References

Abay, A.; Zhou, Y.; Baracaldo, N.; Rajamoni, S.; Chuba, E.; and Ludwig, H. 2020. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*.

Cho, Y. J.; Gupta, S.; Joshi, G.; and Yağan, O. 2020. Bandit-based communication-efficient client selection strategies for federated learning. In *Asilomar Conference on Signals, Systems, and Computers*, 1066–1069.

Du, W.; Xu, D.; Wu, X.; and Tong, H. 2021. Fairness-aware agnostic federated learning. In *SDM*, 181–189.

Dua, D.; and Graff, C. 2017. UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed: 2024-02-18.

EU, E.; et al. 2012. Charter of fundamental rights of the European Union. *The Review of International Affairs*, 63(1147): 109–123.

Ezzeldin, Y. H.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, A. S. 2023. Fairfed: Enabling group fairness in federated learning. In *AAAI*, 7494–7502.

Fioretto, F.; Mak, T. W.; and Van Hentenryck, P. 2020. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In *AAAI*, 630–637.

Fioretto, F.; Van Hentenryck, P.; Mak, T. W.; Tran, C.; Baldo, F.; and Lombardi, M. 2021. Lagrangian duality for constrained deep learning. In *ECML PKDD*, 118–135.

Gálvez, B. R.; Granqvist, F.; van Dalen, R.; and Seigel, M. 2021. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NeurIPS*, 3315–3323.

Hu, Z.; Shaloudegi, K.; Zhang, G.; and Yu, Y. 2022. Federated learning meets multi-objective optimization. *TNSE*, 9(4): 2039–2051.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and robust federated learning through personalization. In *ICML*, 6357–6368.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2019. Fair Resource Allocation in Federated Learning. In *ICLR*.

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

Mishler, A.; Kennedy, E. H.; and Chouldechova, A. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *ACM Conference on Fairness, Accountability, and Transparency*, 386–400.

Mohri, M.; Sivek, G.; and Suresh, A. T. 2019. Agnostic federated learning. In *ICML*, 4615–4625.

Moro, S.; Cortez, P.; and Rita, P. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62: 22–31.

Pessach, D.; and Shmueli, E. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3): 1–44.

ProPublica. 2016. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. Accessed: 2024-02-18.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2021. FairBatch: batch selection for model fairness. In *ICLR*.

Salimi, B.; Rodriguez, L.; Howe, B.; and Suciu, D. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *International Conference on Management of Data*, 793–810.

Su, C.; Yu, G.; Wang, J.; Li, H.; Li, Q.; and Yu, H. 2023. Multi-dimensional Fair Federated Learning. arXiv:2312.05551.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *ICML*, 6373–6382.

Wang, J.; Charles, Z.; Xu, Z.; Joshi, G.; McMahan, H. B.; Al-Shedivat, M.; Andrew, G.; Avestimehr, S.; Daly, K.; Data, D.; et al. 2021a. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*.

Wang, Z.; Fan, X.; Qi, J.; Wen, C.; Wang, C.; and Yu, R. 2021b. Federated learning with fair averaging. In *IJCAI*, 1615–1623.

Wang, Z.; and Tsvetkov, Y. 2021. Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models. In *ICLR*.

Yue, X.; Nouiehed, M.; and Al Kontar, R. 2023. Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science*, 2(1): 10–23.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: mechanisms for fair classification. In *AISTAT*, 962–970.

Zeng, Y.; Chen, H.; and Lee, K. 2021. Improving fairness via federated learning. In *ICLR*.

Zhao, Z.; and Joshi, G. 2022. A dynamic reweighting strategy for fair federated learning. In *ICASSP*, 8772–8776.