Fairness-aware Federated Minimax Optimization with Convergence Guarantee

Gerry Windiarto Mohamad Dunda
Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
gwmdunda@connect.ust.hk

Shenghui Song
Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
eeshsong@ust.hk

***Abstract*—Federated learning (FL) has garnered considerable attention due to its privacy-preserving feature. Nonetheless, the lack of freedom in managing user data can lead to group fairness issues, where models are biased towards sensitive factors such as race or gender. To tackle this issue, this paper proposes a novel algorithm, fair federated averaging with augmented Lagrangian method (FFALM), designed explicitly to address group fairness issues in FL. Specifically, we impose a fairness constraint on the training objective and solve the minimax reformulation of the constrained optimization problem. Then, we derive the theoretical upper bound for the convergence rate of FFALM. The effectiveness of FFALM in improving fairness is shown empirically on CelebA and UTKFace datasets in the presence of severe statistical heterogeneity.**

***Index Terms*—federated learning, group fairness, convergence rate, augmented Lagrangian**

I. Introduction

Federated learning (FL) [1] is a distributed machine learning approach that enables model training on potentially sensitive data from different entities without the necessity for data sharing. This technique is promising in diverse domains such as computer vision (CV) as it can facilitate training of models on a large-scale, diverse set of data while preserving data privacy. However, a direct implementation of existing federated algorithms may violate group fairness [2], which refers to the equitable treatment of different groups in a population. Group fairness is required by law such as in Europe [3], enforcing that the decision making by predictive models does not exhibit bias towards any particular sensitive group, such as race or gender. For example, an AI model used in a hiring process may have been trained on historical data that reflects biased hiring patterns, leading to discriminatory outcomes for underrepresented groups in the workforce. There are more examples [4] that further motivate raising awareness in training fair deep learning models.

The sources of group unfairness or bias mainly come from dataset, which may reflect measurement or historical bias from the annotators, and the training algorithm, which may learn unwanted biased features from such dataset. The aforementioned sources also induce statistical heterogeneity. Under severe heterogeneity, the trained model may have poor task performance and biased model. Finding FL algorithms that are fair and robust to statistical heterogeneity or non-identical and independently distributed (non-iid) data is an arduous task, and currently it is an open problem [5].

This work was supported by a grant from the NSFC/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and National Natural Science Foundation of China (Project No. N_HKUST656/22).

A. Contributions

Since we have little control on the clients data in FL, efforts to mitigate bias from the federated algorithm are important. Therefore, the focus of this paper is to improve group fairness in binary classification tasks, involving binary sensitive attributes, which is common in fairness literature [6]. In particular, we propose a new federated algorithm that is effective in reducing bias while maintaining similar task performance as existing federated algorithms. The main contributions are summarized below.

- We propose a fairness-aware algorithm, fair federated averaging with augmented Lagrangian method (FFALM). Firstly, we formulate a constrained minimization problem on the global loss function satisfying a fairness metric. Inspired by augmented Lagrangian method [7], we solve the problem by leveraging the local training as a sub-solver to find the optimal model parameters given dual iterates. Then, the dual iterates are locally updated given the updated model parameter, and they are aggregated using weighted average.
- We propose a theoretical upper bound for the convergence rate of FFALM over a nonconvex-strongly-concave objective function, which is $\mathcal{O}(\frac{1}{T^{2/3}})$.
- We empirically assess the proposed method on publicly accessible CV datasets (CelebA and UTKFace) containing sensitive attributes (gender and skin color respectively) based on two common fairness metrics: demographic parity difference (DPD) and equal opportunity difference (EOD). Experimental results show that FFALM improves DPD by 10% and EOD by 14% in the attractiveness prediction task, and DPD by 6% and EOD by 5% in the youth prediction task compared to FedAvg under severe data heterogeneity.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00111

II. Related Work

There have been some engaging results in tackling the fairness issues in deep-learning models. In the following, we categorize some prior related works based on how the training is conducted, either centralized or federated.

A. Ensuring fairness in centralized learning

In centralized learning, it is not uncommon to modify the training framework to achieve a suitable degree of group fairness. The authors of [8] decorrelated the input images and the protected group attributes by using adversarial training. Knowledge transfer techniques and multi-classifiers can also be adopted as a debiasing method [9]. Augmenting each image sample with its perturbed version generated from generative models can potentially reduce biases as well [10]. The aforementioned works require additional components to the model, thus increasing the computation cost. This might not be suitable for FL. A possible alternative is to alter the loss function to take into account group fairness. The authors of [11] introduced a loss function obtained from the upper bound of the Lagrangian based on a constrained optimization formulation from a finite hypothesis space perspective.

B. Ensuring fairness in FL

Some prior works considered group fairness in FL. Due to system constraints, most innovations came from the objective formulation to include fairness or more information exchange between the server and the clients. The example for the latter is FairFed [12], where the client coefficients are adaptively adjusted during the aggregation phase based on the deviation of each client's fairness metric from the global average. Along the line of the objective formulation, FCFL [13] proposed a two-stage optimization to solve a multi-objective optimization with fairness constraints, which demands more communication rounds. FPFL [14] utilized differential multipliers optimization method to solve main objective by taking into account fairness as a constraint, which is similar to this work. FedFB [15] adjusted the weight of the local loss function for each sensitive group during the aggregation phase. The objective formulation of FPFL, however, is not smooth, which may hinder the convergence of gradient-based learning. Moreover, the theoretical convergence guarantee is missing for the aforementioned works.

III. Preliminaries

In this section, we introduce some mathematical notations and group fairness notions. After that, we briefly describe the framework of minimax FL.

A. Notations

Throughout this paper, we primarily focus on supervised binary classification tasks with binary sensitive attributes. The dataset is denoted as $\mathcal{D} = X \times Y \times S$ with size $|\mathcal{D}|$ constituting of an input image X, a label $Y = \{0, 1\}$, and a sensitive attribute $S = \{0, 1\}$. We slightly abuse the notation of $\mathcal{D}$, X, Y, and S to represent both the set and the distribution. The datasets can also be partitioned based on sensitive attributes, $\mathcal{D}^{s_0} = X \times Y \times S_0$ and $\mathcal{D}^{s_1} = X \times Y \times S_1$.

Some mathematical notations are stated as follows. $[N]$ denotes $\{1, 2, ..., N\}$, $\|.\|$ represents the ℓ_2-norm, and $\mathbb{1}$ denotes the indicator function. We use $\mathcal{W} \subseteq \mathbb{R}^d$ and Λ to represent the parameter spaces of the model w and an additional learnable training parameter λ, respectively. Denote $f_w : X \to Y$ as a deep-learning model parameterized by w, taking X as an input, and outputting the predicted label Y, and denote $q_w : X \to \mathbb{R}^2$ as the logits of the model, where the first element corresponds to $Y = 0$ and the second element corresponds to $Y = 1$.

B. Group Fairness Metrics

To evaluate the group fairness performance of a machine learning model f, there are various notions based on how likely the model predicts a favorable outcome ($\hat{Y} = f(X) = 1$) for each group. *Demographic parity* (DP) [16] is commonly used for assessing the fairness of the model. f satisfies DP if the model prediction of favorable label is independent of S, i.e.,

$$\mathbb{E}_{X|S=0}[f(X) = 1] = \mathbb{E}_{X|S=1}[f(X) = 1]. \quad (1)$$

Another way to define the notion of group fairness is *accuracy parity* (AP) [17]. To satisfy this notion, f conforms to the following equality

$$\mathbb{E}_{\mathcal{D}^{s_0}}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{E}_{\mathcal{D}^{s_1}}[\mathbb{1}_{f(X) \neq Y}]. \quad (2)$$

In some use cases where the preference of users belonging to a sensitive group is considered, it is amenable to adopt *equal opportunity* (EO) [18] of positive outcomes for each sensitive attribute as a fairness notion, mathematically written as

$$\mathbb{E}_{X|S=0,Y=1}[f(X) = 1] = \mathbb{E}_{X|S=1,Y=1}[f(X) = 1]. \quad (3)$$

In practice, it is difficult to achieve perfect fairness imposed by aforementioned fairness notions. To measure how close f is to satisfy DP, we employ demographic parity difference metric (Δ_{DP}) on favorable label, which is defined as

$$\Delta_{DP} = |\mathbb{E}_{X|S=0}[\mathbb{1}_{f(X)=1}] - \mathbb{E}_{X|S=1}[\mathbb{1}_{f(X)=1}]|. \quad (4)$$

Similarly, the closeness measure to satisfy EO condition, equal opportunity difference (Δ_{EO}) is defined as

$$\Delta_{EO} = |\mathbb{E}_{X|Y=1,S=0}[\mathbb{1}_{f(X)=1}] - \mathbb{E}_{X|Y=1,S=1}[\mathbb{1}_{f(X)=1}]|. \quad (5)$$

These two closeness metrics are commonly used for assessing group fairness in machine learning models. Since only samples rather than the true data distribution are available, the metrics are estimated using the samples.

C. Minimax FL Framework

For a FL system with N clients and one server, the goal is to train a global deep learning model f_w on each client dataset $\mathcal{D}_i$ $(i \in [N])$ without sharing their datasets. In addition, an additional learning parameter (e.g. Lagrangian dual) $\lambda \in \Lambda$ that aids the training can be exchanged between the server and clients, and processed on the clients and the server. Its

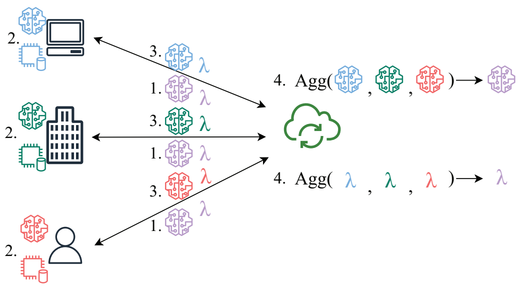


Fig. 1: The minimax FL framework. In each communication round, there are four steps in FL training: 1. Broadcasting phase 2. Local training phase 3. Client-to-server communication phase 4. Aggregation phase.

generic procedure consists of four phases. Firstly, the clients receive the global model from the server (broadcasting phase) and train the model on their own dataset (local training phase). After that, the clients send their model update and the dual parameter update to the server (client-to-server communication phase), and the server aggregates the received updates from each participating client to get an updated model (aggregation phase). This process is repeated until convergence or a specified communication round, as illustrated in Fig. 1.

During local training phase, each client aims to minimize their local risk function $F_i(w, \lambda)$ to update their model. At the same time, it maximizes the local risk function to update λ, in contrast to the conventional FL framework. For any regularization-based federated algorithms, the explicit formulation for the true local risk function of the i−th client represented by a loss function $l(f_w(x), y) = l(x, y; w)$ and a regularization function g is given by

$$F_i(w, \lambda) := \mathbb{E}_{(x_j, y_j) \sim \mathcal{D}_i}[l(x_j, y_j; w)] + g(\mathcal{D}_i; \lambda, w), \quad (6)$$

and the corresponding empirical risk function is given by

$$F_{i,S}(w, \lambda) := \frac{1}{|\mathcal{D}_i|} \sum_{(x_j, y_j) \sim \mathcal{D}_i} l(x_j, y_j; w) + g(\mathcal{D}_i; \lambda, w). \quad (7)$$

The ultimate goal is to solve the following minimax objective of the true global risk function $F(w, \lambda)$

$$\min_w \max_\lambda F(w, \lambda) := \sum_{i=1}^{N} p_i F_i(w, \lambda), \quad (8)$$

where p_i is the client coefficient with $\sum_{i=1}^{N} p_i = 1$ and $p_i \in [0, 1]$. In FedAvg, the coefficient is set to the proportion of the samples from each client. Since the clients only have access to samples rather than the data distribution, the objective is replaced with the global empirical risk function defined as

$$F_S(w, \lambda) := \sum_{i=1}^{N} p_i F_{i,S}(w, \lambda). \quad (9)$$

IV. FFALM

We first introduce the problem formulation for FL with group fairness constraints. Subsequently, we describe the proposed algorithm to achieve the objective. Lastly, we offer the convergence rate of the proposed algorithm.

A. Solving Group Fairness Issue

1) Problem formulation: The objective of this work is to ensure group fairness on the FL-trained binary classification model. We tackle the problem by enforcing fairness during the local training. Specifically, the local training aims to minimize the local risk function while satisfying a notion of fairness. The strategy is to reformulate the local risk function as a sum of the main objective, which is related to the task performance, and the term related to fairness constraint, weighted with a learnable λ. In this way, we can use minimax FL framework as described in the previous section.

One of the essential requirements of minimax FL framework is to have a smooth local risk function. Since the indicator functions appearing in the fairness notions in Section III-B are not differentiable functions with respect to the model parameters, we need to replace all indicator functions with their corresponding surrogate continuous functions. In the case of AP, the choice of such continuous functions is readily available, which is cross entropy loss $CE(y, q_w(x)) = -\log \sigma(q_w(x)_y)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function. This is similar to how 0-1 loss ($\mathbb{E}_{\mathcal{D}}[\mathbb{1}_{\hat{Y} \neq Y}]$) in the basic gradient-based learning can be replaced with cross-entropy loss [19]. The only difference in the formulation of AP is that it is conditioned on each sensitive attribute.

To this end, we write the objective of the local training of the i-th client as a constrained optimization

$$\min_w L_S(w, \mathcal{D}_i) \text{ s.t. } \mu(w, \mathcal{D}_i^{s_0}) = \mu(w, \mathcal{D}_i^{s_1}), \quad (10)$$

where $L_S(w, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{(x_j, y_j) \sim \mathcal{D}} l(x_j, y_j; w)$ and $\mu(w, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} CE(y, q_w(x))$. It is worth mentioning that we also estimate (2) from the samples, as shown in the definition of $\mu(w, \mathcal{D})$.

2) Local training phase: This problem can be approximately solved by following similar techniques from the augmented Lagrangian approach [7] by treating the constraint as a soft constraint. Specifically, it seeks a saddle solution of an augmented Lagrangian function $\mathcal{L}_S$ parameterized by a suitable choice of penalty coefficient (β)

$$\begin{aligned} \mathcal{L}_S(w_t, \lambda_{t-1}, \mathcal{D}_i) &:= L_S(w_t, \mathcal{D}_i) + \frac{\beta}{2} \Delta_\mu(w_t, \mathcal{D}_i)^2 \\ &+ \lambda_{t-1} \Delta_\mu(w_t, \mathcal{D}_i), \end{aligned} \quad (11)$$

where $\Delta_\mu(w_t, \mathcal{D}_i) := \mu(w_t, \mathcal{D}_i^{s_0}) - \mu(w_t, \mathcal{D}_i^{s_1})$ by introducing a sub-optimizer O to seek w_t such that $\|\nabla_w \mathcal{L}_S(w_t, \beta, \mathcal{D}_i)\|$ is sufficiently small. Afterwards, λ_{t-1} is updated to close the infeasibility gap, and the process is repeated. If the algorithm converges to the solution (w^*, λ^*) of (11) that satisfies second-order sufficient conditions [7], w^* is the global solution to (10).

Translating this view into FL, we can assign the sub-optimizer O to the local training and the iteration index t to the communication round. Hence, we can formulate the local training of i-th client as a two-stage process

$$w_{i,t} = \min_w F_{i,S}(w, \lambda_{t-1}) = \min_w \mathcal{L}_S(w, \lambda_{t-1}, \mathcal{D}_i) \quad (12)$$

$$\lambda_{i,t} = \lambda_{t-1} + \eta_{\lambda,t}\Delta_\mu(w_t, \mathcal{D}_i). \quad (13)$$

Note that in the original augmented Lagrangian method, $\eta_{\lambda,t}$ is set to β. As shown later in the experiment results section, this proposed two-stage optimization gives more competitive results in terms of fairness performance.

Stochastic gradient descent is used to solve (12), similar to FedAvg. Specifically, the i-th client computes the stochastic gradient of $\mathcal{L}_i^{(t,k)} = \mathcal{L}_S(w_i^{(t,k)}, \lambda_{i,t}, \mathcal{B}_i)$ at communication round t and local iteration k from its batch samples $\mathcal{B}_i$ sampled from its local distribution $\mathcal{D}_i$ as

$$\nabla_w \mathcal{L}_i^{(t,k)} = \nabla_w \Bigg(L_S(w_i^{(t,k)}, \mathcal{B}_i) + \lambda \Delta_\mu(w_i^{(t,k)}, \mathcal{B}_i) + \frac{\beta}{2}\Delta_\mu(w_i^{(t,k)}, \mathcal{B}_i)^2 \Bigg). \quad (14)$$

3) Aggregation phase: The server receives model updates, $w_{i,t}$, as well as the dual updates, $\lambda_{i,t}$, from the clients. Following FedAvg, the received dual update from each client is aggregated by weighted average with the same client coefficient (p_i) as model aggregation

$$w_t = \sum_{i=1}^N p_i w_{i,t} \quad \text{and} \quad \lambda_t = \sum_{i=1}^N p_i \lambda_{i,t}. \quad (15)$$

The proposed algorithm is summarized in Algorithm 1.

Algorithm 1 FFALM Algorithm

Inputs: N, $\{\mathcal{D}_i\}_{i=1}^N$, β, $\eta_{w,t}$, the number of local iteration E, and the maximum communication round T.
Randomly initialize the global model w_0 and set $\lambda_0 = 0$ on the server side
for $t = 1$ to T **do**
  Broadcast w_{t-1} and λ_{t-1} to all clients
  for each $i \in [N]$ **do**
    $w_i^{(t,0)} \leftarrow w_{t-1}$
    for $k = 1$ to E **do**
      Randomly sample the batch $\mathcal{B}_i$ from $\mathcal{D}_i$
      Compute $\nabla_w \mathcal{L}_i^{(t,k-1)}$ from (14)
      $w_i^{(t,k)} \leftarrow w_i^{(t,k-1)} - \eta_{w,t}\nabla_w \mathcal{L}_i^{(t,k-1)}$
    end for
    Compute $\lambda_{i,t}$ from (13)
    Send $w_{i,t} = w_i^{(t,E)}$ and $\lambda_{i,t}$ to the server
  end for
  Update $\eta_{w,t}$ using a LR scheduler
  Aggregation phase to obtain w_t and λ_t following (15)
end for

B. Theoretical Convergence Guarantee

The proposed algorithm can be viewed as solving a minimax problem with $F_S(w,\lambda) = \sum_{i=1}^N p_i \mathcal{L}_S(w, \lambda_{t-1}, \mathcal{D}_i)$. We provide the upper bound of the convergence rate based on how close the empirical primal risk function $R_S(w_t) := \max_\lambda \mathcal{L}(w_t, \lambda)$ is to the optimal. Before presenting the result, we list several definitions and key assumptions.

Definition 1. *Define a function $h : \mathcal{W} \times \Lambda \to \mathbb{R}$. $h(\cdot,\cdot)$ is L-smooth if it is continuously differentiable and there exists a constant $L > 0$ such that for any $w, w' \in \mathcal{W}$, $\lambda, \lambda' \in \mathbb{R}$, and $\xi \in \mathcal{D}$,*

$$\left\| \begin{pmatrix} \nabla_w h(w,\lambda;\xi) - \nabla_w h(w',\lambda';\xi) \\ \nabla_\lambda h(w,\lambda;\xi) - \nabla_\lambda h(w',\lambda';\xi) \end{pmatrix} \right\| \le L \left\| \begin{pmatrix} w - w' \\ \lambda - \lambda' \end{pmatrix} \right\|.$$

Definition 2. *$h(w,\cdot)$ is ρ-strongly convex if for all $w \in \mathcal{W}$ and $\lambda, \lambda' \in \Lambda$, $h(w,\lambda) \ge h(w,\lambda') + \langle \nabla_\lambda h(w,\lambda'), \lambda - \lambda' \rangle + \frac{\rho}{2}\|\lambda - \lambda'\|^2$.*

Definition 3. *$h(w,\cdot)$ is ρ-strongly concave if $-h(w,\cdot)$ is ρ-strongly convex.*

Assumption 1. *For randomly drawn batch samples ξ and for all $i \in [N]$, the gradients $\nabla_w F_{i,S}(w,\lambda;\xi)$ and $\nabla_\lambda F_{i,S}(w,\lambda;\xi)$ have bounded variances B_w and B_λ respectively. If $g_{i,w}(w,\lambda|\xi) := \nabla_w F_{i,S}(w,\lambda;\xi)$ is the unbiased local estimator of the gradient, $\mathbb{E}_\xi[\|g_{i,w}(w,\lambda|\xi) - \nabla_w F_{i,S}(w,\lambda)\|^2] \le B_w^2$, and the case for λ is similar but bounded by B_λ^2.*

Assumption 2. *For all $i \in [N]$, the stochastic gradient of $F_{i,S}(w,\lambda)$ is bounded by a constant D. Specifically, for all $w \in \mathcal{W}$ and $\lambda \in \Lambda$, we have $\|\nabla_w F_{i,S}(w,\lambda)\| \le D$.*

Definition 4. *A function $h(\cdot,\lambda)$ satisfies the PL condition if for all λ, there exists a constant $\mu > 0$ such that, for any $w \in \mathcal{W}$, $\frac{1}{2}\|\nabla h(w)\|^2 \ge \mu(h(w) - \min_{w' \in \mathcal{W}} h(w'))$.*

Definition 1 and Assumption 1 are commonly used for federated learning [20]. Definition 3 for $h = \mathcal{L}$ is satisfied because it is a linear function. Assumption 2 is satisfied when the gradient clipping method is employed. Lastly, the PL-condition of $\mathcal{L}(\cdot,\lambda)$ is shown to hold on a large class of neural networks [21].

For simplicity, we assume full participation and the same number of local iterations for each client. The minimum empirical primal risk is $R_S^* = \min_w R_S(w)$. The upper bound of the convergence rate of FFALM is given by the following theorem.

Theorem 1. *Define $\kappa = \frac{L}{\mu}$. Let $\eta_{w,t} = \frac{2}{\mu t}$ and $\eta_{\lambda,t} = \frac{16\kappa^2}{\mu t^{2/3}}$. Given that Assumption 1 and Assumption 2 hold, each $F_{i,S}(w,\lambda)$ is L-smooth, each $F_{i,S}(\cdot,\lambda)$ satisfies μ-PL condition, and each $F_{i,S}(w,\cdot)$ is ρ-strongly concave, we have*

$$\mathbb{E}R_S(w_{T+1}) - R_S^* = \mathcal{O}\left(\frac{\Gamma + B_w^2 + B_\lambda^2}{T^{2/3}}\right),$$

after T communication rounds, where $\Gamma := F_S^* - \sum_{i=1}^{N} p_i F_{i,S}^*$, $F_S^* := \min_w \max_\lambda F_S(w, \lambda)$ *and* $F_{i,S}^* := \min_w \max_\lambda F_{i,S}(w, \lambda)$.

Proof: Due to space constraints, the proof is omitted, and can be found in the Arxiv version [22]. ■

Note that Γ quantifies statistical heterogeneity of the FL system. In the case of strong non-iid, the saddle solution of the global risk function is significantly different from the weighted sum of each saddle local risks.

V. Empirical Study

In this section, we evaluate the effectiveness of FFALM based on three important performance metrics: the prediction accuracy, DPD, and EOD on real-world datasets. We provide the results and comparison with other existing FL algorithms.

A. Datasets and Tasks

We want to investigate how the classification model trained in FL, which is designed to further increase the prediction accuracy on specific domain, influences the fairness performance. In particular, two datasets from CV domain are used in this study: CelebA [23], and UTKFace [24], and ResNet-18 [25] models are used for both datasets. The task of CelebA dataset is a binary classification for predicting attractiveness in images with gender as the sensitive attribute, and the task of UTKFace dataset is to determine whether the age of a person from an image is above or below 20 with skin color as the sensitive attribute. The favorable labels for CelebA and UTKFace are attractive and age below 20 respectively. Each dataset is split into three categories: training set, validation set, and testing set.

B. FL Setting

There are 10 clients participating in FL training. We synthetically simulate statistical heterogeneity by introducing label skews, which can be implemented using Dirichlet distribution parameterized by α on the proportion of samples for a given class and client on a given centralized samples [26]. The case of severe data heterogeneity is investigated in this experiment by setting $\alpha = 0.3$. Each experiment is repeated 10 times with different seeds. The FL training ends after 70 communication rounds.

C. Baselines

The following are the baselines used for the comparison study.

1) **FedAvg.** It is the universal baseline in FL which trains the model locally without considering fairness and aggregates all model updates by weighted average.
2) **FairFed [12].** The server receives the local DP metrics, and based on them and the global trend, the server adjusts the value of p_i adaptively before averaging the model updates.
3) **FPFL [14].** It enforces fairness by solving the constrained optimization on the sample loss function L_S with two constraints. These constraints ensure that the absolute difference between the overall loss and the loss of each sensitive group ($\delta_{\mu_0,i} = [L(w^t, \mathcal{D}_i) - \mu(w^t, \mathcal{D}_i^{s_0})]_+$ and $\delta_{\mu_1,i} = [L(w^t, \mathcal{D}_i) - \mu(w^t, \mathcal{D}_i^{s_1})]_+$, where $[x]_+ := \max(0, x)$) does not deviate from a specific threshold. We set this threshold to be zero. Hence, we reformulate it as a local constrained optimization with

$$g(\mathcal{D}_i; \lambda, w) = \lambda_0 \delta_{\mu_0,i} + \lambda_1 \delta_{\mu_1,i} + \frac{\beta}{2}(\delta_{\mu_0,i}^2 + \delta_{\mu_1,i}^2). \tag{16}$$

TABLE I: Comprehensive list of hyperparameter values used in the experiments on CelebA and UTKFace datasets for baselines and FFALM.

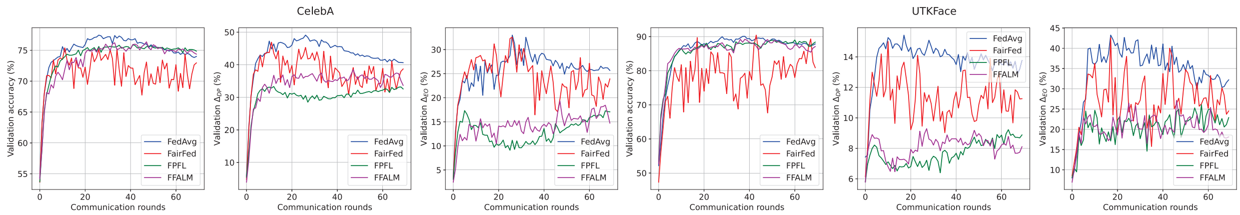
Hyperparameters	Algorithms	Values
Batch size	all	128
Gradient clipping on w	all	1.0
LR of w decay step size	all	50
LR or w decay step factor	all	0.5
$\eta_{w,0}$	all	0.05
β	FairFed	0.5
β	FPFL	5.0
$\eta_{\lambda,t}$	FPFL	0.5
b	FFALM	1.05
β	FFALM	2.0
$\eta_{\lambda,0}$	FFALM	2.0
λ_0	FFALM and FPFL	0.0

D. Implementation Details

The hyperparameters used in the experiment for FFALM and all baselines are shown in Table I. Following the setup from augmented Lagrangian method, we slowly increase the learning rate of λ by a factor b per communication round for FFALM. All algorithms have the same learning rate scheduler of w which is realized by a constant-step decay factor.

TABLE II: Comparison of the performance of the proposed algorithm across baselines on CelebA and UTKFace datasets. ↑ indicates the larger the value the better and ↓ indicates the smaller the value the better.

Algorithm	Acc ↑ (%)	Δ_{DP} ↓ (%)	Δ_{EO} ↓ (%)
	CelebA		
FedAvg	**74.66**	39.54	21.11
FairFed	73.27	37.36	19.50
FPFL	74.58	30.87	9.71
FFALM	74.06	**28.92**	**6.82**
	UTKFace		
FedAvg	86.22	13.07	19.12
FairFed	79.54	10.80	17.92
FPFL	**86.73**	7.97	15.85
FFALM	86.02	**6.60**	**13.87**



(a) (b)

Fig. 2: Learning curves on validation set for (a) CelebA dataset and (b) UTKFace dataset

E. Results

The training curves of all algorithms are shown in Figure 2. The curves show how accuracy, DPD, and EOD change as the communication round increases. It can be observed that FFALM shows similar improvement in accuracy as FedAvg while improving the fairness performance.

The experimental results evaluated on the testing set are presented in Table II. For celebA dataset, FFALM improves DPD by almost 11% and EOD by roughly 14% compared to FedAvg. FFALM outperforms other baselines in fairness performance with minimal accuracy loss. For UTKFace dataset, FFALM improves DP difference performance by about 6%, and reduces EO gap by 6% compared to FedAvg. The overall fairness improvement is also apparent on FFALM in UTKFace dataset compared with different baselines.

VI. Conclusion

In this paper, we proposed FFALM, an FL algorithm based on augmented Lagrangian framework to handle group fairness issues. It leveraged accuracy parity constraint for smooth loss formulation of minimax FL framework. It was shown that the theoretical convergence rate of FFALM is $\mathcal{O}(\frac{1}{T^{2/3}})$. Experiment results on CelebA and UTKFace datasets demonstrated the effectiveness of the proposed algorithm in improving fairness with negligible accuracy drop under severe statistical heterogeneity.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on AISTATS* (A. Singh and J. Zhu, eds.), vol. 54, pp. 1273–1282, PMLR, 2017.

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th SIGKDD*, ACM, 2008.

[3] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council."

[4] A. Howard and J. Borenstein, "The ugly truth about ourselves and our robot creations: The problem of bias and social inequity," *Science and Engineering Ethics*, vol. 24, no. 5, pp. 1521–1536, 2017.

[5] P. Kairouz and et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, p. 1–210, 2021.

[6] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, aug 2023.

[7] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2e ed., 2006.

[8] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *ICCV*, 2019.

[9] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv:1405.6245*, 2017.

[10] V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *IEEE/CVF Conference on CVPR*, 2021.

[11] V. S. Lokhande, A. K. Akash, S. N. Ravi, and V. Singh, "FairALM: Augmented lagrangian method for training fair models with little regret," in *Computer Vision – ECCV 2020*, pp. 365–381, Springer International Publishing, 2020.

[12] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and A. S. Avestimehr, "FairFed: Enabling group fairness in federated learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 7494–7502, 2023.

[13] S. Cui, W. Pan, J. Liang, C. Zhang, and F. Wang, "Addressing algorithmic disparity and performance inconsistency in federated learning," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 26091–26102, Curran Associates, Inc., 2021.

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

[15] Y. Zeng, H. Chen, and K. Lee, "Improving fairness via federated learning," *arXiv:2110.15545*, 2022.

[16] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[17] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, (Republic and Canton of Geneva, CHE), p. 1171–1180, International World Wide Web Conferences Steering Committee, 2017.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016.

[19] M. I. J. Peter L Bartlett and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[20] J. Wang and et al., "A field guide to federated optimization," *arXiv:2107.06917*, 2021.

[21] C. Liu, L. Zhu, and M. Belkin, "Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning," *arXiv:2003.00307*, 2020.

[22] G. W. M. Dunda and S. Song, "Fairness-aware federated minimax optimization with convergence guarantee," *arXiv:2307.04417*, 2024.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of ICCV*, 2015.

[24] Z. Zhifei, S. Yang, and Q. Hairong, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on CVPR*, IEEE, 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.

[26] H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv:1909.06335*, 2019.