PROVABLY FAIR FEDERATED LEARNING

Shengyuan Hu
CMU
shengyua@andrew.cmu.edu

Zhiwei Steven Wu
CMU
zstevenwu@cmu.edu

Virginia Smith
CMU
smithv@cmu.edu

ABSTRACT

In federated learning, fair prediction across various protected groups (e.g., gender, race) is an important constraint for many applications. Unfortunately, prior work studying group fair federated learning lacks formal convergence or fairness guarantees. Our work provides a new definition for group fairness in federated learning based on the notion of Bounded Group Loss (BGL), which can be easily applied to common federated learning objectives. Based on our definition, we propose a scalable algorithm that optimizes the empirical risk and global fairness constraints, which we evaluate across common fairness and federated learning benchmarks. Our resulting method and analysis are the first we are aware of to provide formal theoretical guarantees for training a fair federated learning model.

1 INTRODUCTION

Federated learning (FL) is a training paradigm that aims to fit a model to data generated by, and residing in, a set of disparate data silos, such as a network of remote devices or collection of organizations (McMahan et al., 2017). Mirroring concerns around fairness in non-federated settings, many FL applications similarly require performing fair prediction across protected groups. However, naively estimating algorithmic fairness locally for each silo in a federated network is inaccurate due to heterogeneity across silos—failing to produce a fair model over the entire dataset (Zeng et al., 2021). To address this, several recent works have aimed to implement notions of group fairness in federated networks (Chu et al., 2021; Zeng et al., 2021; Papadaki et al., 2022). Unfortunately, despite their promising empirical performance, these prior works are heuristic in nature in that they lack any guarantees surrounding the resulting fairness of the solutions. In this work we provide the first method we are aware for group fairness in federated learning that comes with formal convergence and fairness guarantees. In developing and analyzing our approach, we also take care to ensure that the proposed method addresses practical constraints of realistic federated networks—building off common communication-efficient federated optimization methods which can scale to networks of millions of devices. We demonstrate the effectiveness of our theoretically-grounded approach on common benchmarks from fair machine learning and federated learning.

The remainder of the paper is organized as follows. In Section 2, we formalize the *Bounded Group Loss* fairness definition and corresponding fair federated learning objective. In Section 3, we present a scalable algorithm to solve the proposed objective, and provide formal convergence and fairness guarantees for our objective and algorithm. In Section 4, we evaluate our algorithm on common fairness benchmark and show our method is able to achieve both better utility and fairness performance compared to vanilla FedAvg. We defer a detailed discussion of related work to Appendix B.

2 FAIR FEDERATED LEARNING SETUP

Following standard federated learning scenarios (McMahan et al., 2017), we consider a network with K different clients. Each client $k \in [K]$ has access to data $\hat{\mathcal{D}}_k := \{(x_i, y_i, a_i)\}_{i=1,\cdots,m_k}$ sampled from the true data distribution $\mathcal{D}_k$, where x_i is an observation, y_i is the label, a_i is the protected attribute. Let the hypothesis class be $\mathcal{H}$ and for any model $h \in \mathcal{H}$, define the loss function on data (x, y, a) to be $l(h(x), y)$. Federated learning applications typically aim to solve:

$$\min_{h\in\mathcal{H}} F(h) = \min_{h\in\mathcal{H}} \frac{1}{K}\sum_{k=1}^{K} \frac{1}{m_k}\sum_{i=1}^{m_k} l(h(x_{k,i}), y_{k,i})\,. \tag{1}$$

For simplicity, we define $f_k(h) = \frac{1}{m_k}\sum_{i=1}^{m_k} l(h(x_{k,i}), y_{k,i})$ as the local objective for client k. Further, we assume h is parameterized by a vector $w \in \mathbb{R}^p$ where p is the number of parameters. We will use $F(w)$ and $f_k(w)$ to represent $F(h)$ and $f_k(h)$ in the remainder of the paper.

To learn a model that satisfies any fairness constraint, a standard approach would be solve:

$$\min_{h\in\mathcal{H}} F(h) \quad \text{subject to } \mathbf{R}(h), \tag{2}$$

where $\mathbf{R}(h)$ represents a constraint set on h. Prior works (Zeng et al., 2021; Chu et al., 2021) studying group fairness in federated learning proposed choosing a bounded group-specific parity difference for $\mathbf{R}(h)$. In this work, we focus on a different fairness definition known as *Bounded Group Loss (BGL)* (Agarwal et al., 2019) (defined below). We discuss the motivation of using BGL in Section 2.1

Definition 1. *A classifier h satisfies Bounded Group Loss(BGL) at level ζ under distribution $\mathcal{D}$ if for all $a \in A$, we have*

$$\mathbb{E}_{(x,y,a)\sim\mathcal{D}}\left[l(h(x), y)|A = a\right] \leq \zeta\,. \tag{3}$$

In practice, we could define empirical bounded group loss constraint at level ζ under the empirical distribution $\widehat{\mathcal{D}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathcal{D}}_k$ to be

$$\frac{1}{m_a}\sum_{k=1}^{K}\sum_{a_{k,i}=a} l(h(x_{k,i}), y_{k,i}) \leq \zeta\,. \tag{4}$$

In the rest of the paper, we will refer problem 2 as the constrained optimization problem with $\mathbf{R}(h)$ replaced by Equation 4, which is the main problem we propose to solve.

2.1 Fairness-aware objective

A common method to solve the constrained optimization problem (2) is to use Lagrangian multipliers. This converts the objective into the following saddle point optimization problem:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|A|}, \|\boldsymbol{\lambda}\|_1\leq B}\ \min_{w} G(w;\boldsymbol{\lambda}) = F(w) + \boldsymbol{\lambda}^T\mathbf{r}(w)\,, \tag{5}$$

where the a-th index of $\mathbf{r}$ is $\frac{1}{m_a}\sum_{k=1}^{K}\sum_{a_{k,i}=a} l(h(x_{k,i}), y_{k,i}) - \zeta_a$.

Why use BGL rather than another fairness constraint? Many prior works choose the gap between every two group's loss as the fairness constraint and optimize the Lagrangian. Under such settings, the objective becomes non-convex in terms of the model weight, making it likely that a solver will find a local minima that either does not satisfy the fairness constraint or achieves poor utility. Different from these approaches, BGL requires that for each group $a \in A$, the classifier h's loss evaluated on all data with protected attribute a is below a certain threshold. Therefore, given that the empirical risk is convex, adding the BGL constraint preserves convexity. As we will see, a major benefit of using BGL relative to other alternatives is that it can satisfy meaningful fairness constraints while preserving convexity, enabling both strong empirical performance and formal theoretical guarantees.

3 Provably Fair Federated Learning via Bounded Group Loss

In this section, we first provide a scalable solver for Equation 5 in Algorithm 1. We then provide both a formal convergence and fairness guarantee for our approach.

3.1 Algorithm

To find a saddle point for Objective 5, we follow the scheme from Freund & Schapire (1997) and summarize our solver for the fair FL with bounded group loss in Algorithm 1. Our algorithm is based off of FedAvg (McMahan et al., 2017), a common scalable federated optimization method. Our method alternates between two steps: (1) given a fixed $\boldsymbol{\lambda}$, find w that minimizes $F(w) + \boldsymbol{\lambda}^T\mathbf{r}$; (2) given a fixed w, find $\boldsymbol{\lambda}$ that maximizes $\boldsymbol{\lambda}^T\mathbf{r}$. In Algorithm 1, we provide an example in which the first step is achieved by using FedAvg to solve $\min_w F(w) + \boldsymbol{\lambda}^T\mathbf{r}$ (line 4-10). Note that solving this objective does not require the FedAvg solver; any algorithm that learns a global model in federated learning could be used to find a certain w given $\boldsymbol{\lambda}$. Following Algorithm 2 in Agarwal et al. (2019), we use exponentiated gradient descent to update λ after training a federated model for each round.

Note that the ultimate goal to solve for Objective 5 is to find a w such that it minimizes the empirical risk subject to $\mathbf{r}(w) \leq 0$. Therefore, at the end of training, our algorithm checks whether the resulting model $\bar{w}$ violates the fairness guarantee by at most some constant error $\frac{M+2\nu}{B}$ where M is the upper bound for the empirical risk and ν is the upper bound provided in Equation 7 (line 16-20). We will show in the Lemma 6 that this is always true when there exists a solution w^* for Problem 2. However, it is also worth noting that the Problem 2 does not always have a solution w^*. For example when we set $\zeta = 0$, requiring $\mathbf{r}(w) \leq 0$ is equivalent to requiring the empirical risk given any group $a \in A$ is non positive, which is only feasible when the loss is 0 for every data in the dataset. In this case, our algorithm will simply output *null* if the fairness guarantee is violated by an error larger than $\frac{M+2\nu}{B}$.

3.2 Convergence guarantee

In this section we provide a formal convergence guarantee for Algorithm 1 in solving the empirical risk objective $G(\cdot;\cdot)$. Note that G is linear in $\boldsymbol{\lambda}$. Hence, given a fixed w_0, we can find a solution to the problem $\max_{\boldsymbol{\lambda}} G(w_0;\boldsymbol{\lambda})$, denoted as $\boldsymbol{\lambda}^*$, i.e. $G(w_0;\boldsymbol{\lambda}^*) \geq G(w_0;\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$. When G is convex in w, we can argue that given a fixed $\boldsymbol{\lambda}_0$, there exists w^* that satisfies $w^* = \arg\min_w G(w;\boldsymbol{\lambda}_0)$, i.e. $G(w^*;\boldsymbol{\lambda}_0) \leq G(w;\boldsymbol{\lambda}_0)$ for all w. Therefore, $(w^*,\boldsymbol{\lambda}^*)$ is a saddle point of $G(\cdot;\cdot)$.

To show how the solution found by our algorithm compares to an actual saddle point of G, we introduce the notion of a ν-approximate saddle point.

Definition 2. *$(\widehat{w},\widehat{\boldsymbol{\lambda}})$ is a ν-approximate saddle point of G if*

$$\begin{aligned} G(\widehat{w},\widehat{\boldsymbol{\lambda}}) &\leq G(w,\widehat{\boldsymbol{\lambda}}) + \nu \quad \textit{for all } w \\ G(\widehat{w},\widehat{\boldsymbol{\lambda}}) &\geq G(\widehat{w},\boldsymbol{\lambda}) - \nu \quad \textit{for all } \boldsymbol{\lambda} \end{aligned} \tag{6}$$

As an example, $(w^*,\boldsymbol{\lambda}^*)$ is a 0-approximate saddle point of G. Now we present our main theorem of convergence below.

Theorem 1. *Let Assumption 1-3 hold. Define $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, J\}$ and the learning rate $\eta_Q = \frac{2}{(1+B)\mu(\gamma+t)}$, and assume $\|\mathbf{r}\|_\infty \leq \rho$. Letting $\bar{w} = \frac{1}{ET}\sum_{t=1}^{ET} w^t$, $\bar{\boldsymbol{\lambda}} = \frac{1}{ET}\sum_{t=1}^{ET} \boldsymbol{\lambda}^t$, we have*

$$\max_{\boldsymbol{\lambda}} G(\bar{w};\boldsymbol{\lambda}) - \min_w G(w;\bar{\boldsymbol{\lambda}}) \leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1} C + \frac{B\log(|A|+1)}{\eta_\theta ET} + \eta_\theta \rho^2 B \tag{7}$$

where C is a constant.

We provide detailed descriptions of assumptions in the appendix. The upper bound in Equation 7 consists of two parts: the error for the FedAvg process to obtain $\bar{w}$ and the error for the Exponentiated Gradient Ascent process to obtain $\bar{\boldsymbol{\lambda}}$. Following Theorem 1, we also provide the following corollary expressing the solution of Algorithm 1 as a ν-approximate saddle point of G:

Corollary 2. *Let $\eta_\theta = \frac{\nu}{2\rho^2 B}$ and $T \geq \frac{1}{\nu(\gamma+1)-2\kappa\mathcal{C}}\left(\frac{4\rho^2B^2\log(|A|+1)(\gamma+1)}{\nu E} + 2\kappa\mathcal{C}(\gamma-1)\right)$, then $(\bar{w},\bar{\boldsymbol{\lambda}})$ is a ν-approximate saddle point of G.*

We provide detailed proofs for both Theorem 1 and Corollary 2 in the appendix.

3.3 Fairness guarantee

In the previous section, we demonstrated that our Algorithm 1 converges to a ν-approximate saddle point of the objective G. In this section, we further motivate why we care about finding a ν-approximate saddle point. Eventually, the model learned will be evaluated on test data and data from silo not seen during training. Define the true data distribution to be $\mathcal{D} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{D}_k$. We would like to formalize how well our model is evaluated on the true distribution $\mathcal{D}$ as well as how well the fairness constraint is satisfied under $\mathcal{D}$. The result is presented below in Theorem 3.

Theorem 3. *Let Assumption 1-4 holds. Let $\mathcal{F}$ be the expected risk over the true distribution $\mathcal{D}$, $(\bar{w},\bar{\boldsymbol{\lambda}})$ be a ν-approximate saddle point of G. Then with probability $1-\delta$, either there doesn't exist solution for Equation 2 and Algorithm 1 returns null or Algorithm 1 returns $\bar{w}$ satisfies*

$$\begin{aligned} \mathcal{F}(\bar{w}) &\leq \mathcal{F}(w^*) + 2\nu + 4\mathfrak{R}_m(\mathcal{H}) + \tfrac{2M}{K}\sqrt{\textstyle\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)}, \\ \mathfrak{r}_a(\bar{w}) &\leq \tfrac{M+2\nu}{B} + 2\mathfrak{R}_a(\mathcal{H}) + \tfrac{M}{m_a}\sqrt{\tfrac{K}{2}\log(1/\delta)} \end{aligned} \tag{8}$$

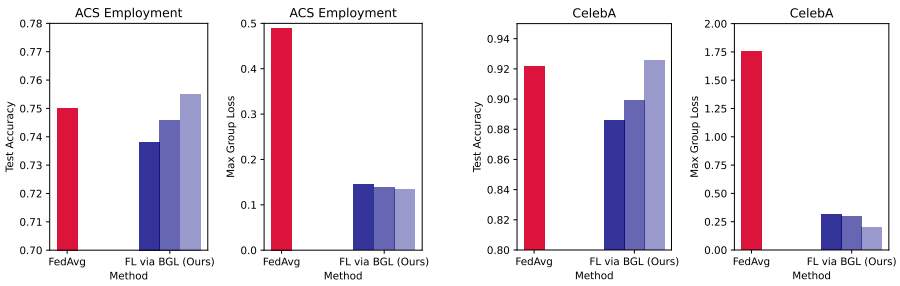

Figure 1: Comparison between FedAvg and our FL via BGL in terms of test accuracy and max group loss on ACS Employment (Left) and CelebA (Right). For each dataset, we select three models for our method to show the relation between test accuracy and fairness (max group loss).

where w^* is a solution for Equation 2, $\mathfrak{r}_a = \mathbb{E}_{(x,y,a)\sim\mathcal{D}}\left[l(h(x),y)|A=a\right] - \zeta_a$.

Note that the fairness constraint for group a under true distribution in Equation 8 is upper bounded by $\mathcal{O}\left(\frac{\sqrt{K}}{m_a}\right)$. For any group a_0 with sufficient data, i.e. m_{a_0} is large, the BGL constraint with respect to group a_0 under $\mathcal{D}$ has stronger formal fairness guarantee compared to any group with less data. We could also see that as the number of silo increases, the upper bound becomes weaker. We provide details and proof for Theorem 3 in the appendix.

4 Experiments

In this section, we evaluate our Algorithm 1 on US-wide ACS PUMS data, a recent group fairness benchmark dataset and CelebA (Caldas et al., 2018), a common federated learning dataset. We compare our method with training a vanilla FedAvg model in terms of both fairness and utility (Section 4.1). We further show the empirical difference between training with our global BGL constraint vs. local BGL constraints in Appendix F.

Setup. For all experiments, we evaluate the accuracy and the empirical loss for each group on test data that belongs to all the silos of our fair federated learning solver. We consider the ACS Employment task (Ding et al., 2021) with race as a protected attribute and CelebA (Caldas et al., 2018) with gender as a protected attribute. A detailed description of datasets and models can be found in the appendix.

4.1 Fairness-Utility Relationship of Algorithm 1

We first explore how test accuracy differs as a function maximum group loss using our Algorithm 1. To be consistent with our method and theoretical analysis, we exclude the protected attribute a_i for each data as a feature for learning the predictor. For each dataset, given a fixed number of training iterations E and T, we finetune B and ζ and evaluate both test accuracy and test loss on each group. Given a certain test accuracy, we select the hyperparameter pair (B, ζ) that yields the lowest maximum group loss. We show the relation between test accuracy vs. max group loss in Figure 2 (Left). On both datasets, our method not only yields a model with significantly smaller maximum group loss than vanilla FedAvg, but also achieves higher test accuracy than the baseline FedAvg which is unaware of group fairness. Therefore, our method yields a model where utility can coexist with fairness constraints relying on Bounded Group Loss.

5 Conclusion and Future Work

In this work, we propose a fair learning objective for federated settings via Bounded Group Loss. We then propose a scalable algorithm to find an approximate saddle point for the objective. Theoretically, we provide convergence and fairness guarantees for our method. Empirically, we show that on ACS Employment and CelebA tasks, our method satisfies high accuracy and strong fairness simultaneously. We are interested in further empirically evaluating our approach in future work, as well as characterizing the difference between using local BGL and global BGL from a theoretical perspective. As our method focuses on optimizing a different fairness constraint (BGL) compared to prior works, we would also like to explore connections between BGL and other fairness notions and expand our framework to cover additional fairness settings.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 2019.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings, `https://leaf.cmu.edu/`. *arXiv preprint arXiv:1812.01097*, 2018.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 2017.

Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 2021.

Kate Donahue and Jon Kleinberg. Models of fairness in federated learning. *arXiv preprint arXiv:2112.00818*, 2021.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*. PMLR, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 2018.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*. PMLR, 2019.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. *arXiv preprint arXiv:2201.08304*, 2022.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*. PMLR, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*. PMLR, 2013.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

Table 1

Dataset	Number of Silos	Model	Protected Attribute	Partition Type	Task Type
ACS Employment (Ding et al., 2021)	50	Logistic Regression	Race	Natural partition by States	Binary classification
CelebA (Liu et al., 2015; Caldas et al., 2018)	50	4-layer CNN	Gender	Manual partition	Binary classification

A Datasets and Models

We summarize the details of the datasets and models we used in our empirical study in Table 1. Our experiments include both convex (Logistic Regression) and non-convex (CNN) loss objectives on both fairness (ACS Employment) and federated learning (CelebA) benchmarks.

B Background and Related Work

Algorithmic Fairness in Machine Learning. Algorithmic fairness in machine learning literature often refers to protection of a protected attribute during the process of learning a model. Three common family of approaches to obtain fairness are pre-processing methods that modify the input data (e.g., Zemel et al., 2013; Feldman et al., 2015; Calmon et al., 2017); post-processing methods that revise the prediction score (e.g., Hardt et al., 2016; Dwork et al., 2018; Menon & Williamson, 2018); and training methods that optimize an objective with some fairness constraints (e.g., Agarwal et al., 2018; 2019; Zafar et al., 2017a;b; Woodworth et al., 2017). All of these methods are based on using a centralized dataset to train and evaluate a model. In our setting where data is privately distributed across different data silos, directly applying these methods is not applicable in order to achieve global fairness across all silos.

Fair Federated Learning. In federated learning, fairness could refer to multiple definitions. One commonly used notion is representation parity (Hashimoto et al., 2018) whose application in FL requires the model's performance across all devices to have small variance. There are a line of recent works that study this notion of fairness in the context of federated learning (Mohri et al., 2019; Li et al., 2019a; Donahue & Kleinberg, 2021). In this work we aims at achieving algorithmic fairness in federated learning, where every data belongs to a specific protected group. The purpose of learning is to find a model that doesn't introduce bias towards any group. In the rest of the work we use the word *fair(ness)* to represent the notion of *group fair(ness)*. Recent works have proposed various objectives for learning an algorithmic fair model under the federated setting. Zeng et al. (2021) proposed a bi-level optimization objective that minimizes the difference between each group's loss while finding an optimal global model. Chu et al. (2021) proposed a similar constrained optimization problem by finding the best model subject to an upper bound on the group loss difference. Different from either approach, our method focuses on fairness a constraint based on upperbounding the loss of each group rather than the loss difference between any two groups. Unlike prior work, our work provides formal convergence and fairness guarantees with respect to our algorithm.

C Detailed description of Algorithm 1

We formally introduce our proposed algorithm below.

D Proof of Theorem 1

We first introduce a few assumptions needed for Theorem 1.

Assumption 1. *Let f_k be $\mu-$strongly convex and $L-$smooth for all $k = 1, \cdots, K$.*

Assumption 2. *Assume the stochastic gradient of f_k has bounded variance: $\mathbb{E}[\|\nabla f_i(w_t^k; \xi_t^k) - \nabla f_k(w_t^k)\|^2] \leq \sigma_k^2$ for all $k = 1, \cdots, K$.*

Assumption 3. *Assume the stochastic gradient of f_k is uniformly bounded: $\mathbb{E}[\|\nabla f_k(w_t^k; \xi_t^k)\|^2] \leq G^2$ for all $k = 1, \cdots, K$.*

Assumption 4. *Let $\mathcal{F}$ and F be upper bounded by constant M.*

Algorithm 1 FedAvg with BGL

Input: T, $\theta^0 = \mathbf{0}$, η_w, η_θ, w_0, $\bar{w} = \mathbf{0}$, M, ν, B
for $i = 1, \cdots, E$ **do**
  Set
$$\lambda_a = B\frac{\exp(\theta_a^i)}{1 + \sum_{a'} \exp(\theta_{a'}^i)}$$
  for $t = 0, \cdots, T-1$ **do**
    Server broadcasts w^t to all the clients.
    for all k in parallel **do**
      Each task updates its weight w_k for some J iterations
$$w_k^{t+1} = w^t - \eta_w \left(\nabla_{w^t} \left(f_k(w^t) + \boldsymbol{\lambda}^T \mathbf{r}\right)\right)$$
      Each client sends $g_k^{t+1} = w_k^{t+1} - w_k^t$ and $\theta_{a,k}^t$ back to the server.
    end for
    Server aggregates the weight
$$w^{t+1} = w^t + \frac{1}{K}\sum_{k=1}^{K} g_k^{t+1}$$
    Update $\bar{w} = \sum_{t=1}^{T} w^t$ and set $w^0 = w^T$
  end for
  Server updates
$$\theta^{(i+1)} = \theta^i + \eta_\theta \mathbf{r}$$
end for
Server updates $\bar{w} \leftarrow \frac{1}{ET}\bar{w}$
if $\max_a \mathbf{r}_a \leq \frac{M+2\nu}{B}$ **then**
  return $\bar{w}$
else
  return null
end if

Lemma 1 (Li et al. (2019b)). *Let* $\Gamma = F^* - \sum_i p_i F_i^*$, $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, J\}$ *and the learning rate* $\eta_t = \frac{2}{\mu(\gamma+t)}$. *Then FedAvg with full device participation satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} F(w^t) - F^* \leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma + t - 1}\left(\frac{2C}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}[\|w^1 - w^*\|^2]\right)$$

where

$$C = \sum_{i=1}^{N} p_i^2 \sigma_i^2 + 6L\Gamma + 8(J-1)^2 G^2$$

Proof for Theorem 1. Let $m_{a,k}$ be the number of data with protected attribute a for client k. By Assumption 1, we have G_i be $(1 + \sum_a \boldsymbol{\lambda}_a \frac{m_{a,k}}{m_a})\mu$-strongly convex and $(1 + \sum_a \boldsymbol{\lambda}_a \frac{m_{a,k}}{m_a})L$-smooth. Since $\|\lambda\|_1 \leq B$, we have G_i be $(1+B)\mu$-strongly convex and $(1+B)L$-smooth. We first present

the regret bound for w^t

$$\frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\lambda^t) - \min_w \frac{1}{ET}\sum_{t=1}^{ET} G(w;\lambda^t) = \frac{1}{ET}\left(\sum_{t=1}^{ET} G(w^t;\lambda^t) - \min_w \sum_{t=1}^{ET} G(w;\lambda^t)\right) \tag{9}$$

$$= \frac{1}{ET}\left(\sum_{i=0}^{E-1}\sum_{t=1}^{T} G(w^{iT+t};\lambda^i) - \min_w \sum_{t=1}^{ET} G(w;\lambda^t)\right) \tag{10}$$

$$\leq \frac{1}{ET}\left(\sum_{i=0}^{E-1}\left(\sum_{t=1}^{T} G(w^{iT+t};\lambda^i) - \min_w \sum_{t=1}^{T} G(w;\lambda^i)\right)\right) \tag{11}$$

$$= \frac{1}{E}\sum_{i=0}^{E-1}\left(\frac{1}{T}\sum_{t=1}^{T} G(w^t;\lambda^i) - G^*(\lambda^i)\right) \tag{12}$$

$$\leq \frac{1}{ET}\sum_{i=0}^{E-1}\sum_{t=1}^{T}\frac{\kappa}{\gamma+t-1}\left(\frac{2C_i}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{13}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\frac{\kappa}{\gamma+t-1}\left(\frac{2\max_i C_i}{\mu} + \frac{\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{14}$$

Now we present the regret bound for $\boldsymbol{\lambda}^t \in \mathbb{R}_+^{|A|}$. For any $\boldsymbol{\lambda}^t$, let's define $\widetilde{\boldsymbol{\lambda}}^t \in \mathbb{R}_+^{|A|+1}$ such that $\widetilde{\boldsymbol{\lambda}}^t$ satisfies $\|\widetilde{\boldsymbol{\lambda}}^t\|_1 = B$ and the first $|A|$ entries of $\widetilde{\boldsymbol{\lambda}}^t$ is the same as $\boldsymbol{\lambda}^t$. Let $\widetilde{\mathbf{r}}^t \in \mathbb{R}^{|A|+1}$ such that the first $|A|$ entries of $\widetilde{\mathbf{r}}^t$ is the same as $\mathbf{r}^t$ and the last entry of $\widetilde{\mathbf{r}}^t$ is 0. Therefore, we have

$$\boldsymbol{\lambda}^T\mathbf{r}^t = \widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{r}}^t \tag{15}$$

for all λ.

By Shalev-Shwartz et al. (2011), for any $\widetilde{\boldsymbol{\lambda}}$, we have

$$\sum_{t=1}^{ET}\widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{r}}^t \leq \sum_{t=1}^{ET}(\widetilde{\boldsymbol{\lambda}}^t)^T\widetilde{\mathbf{r}}^t + \frac{B\log(|A|+1)}{\eta_\theta} + \eta_\theta\rho^2 BET \tag{16}$$

$$= \sum_{t=1}^{ET}(\boldsymbol{\lambda}^t)^T\mathbf{r}^t + \frac{B\log(|A|+1)}{\eta_\theta} + \eta_\theta\rho^2 BET \tag{17}$$

Therefore, we have

$$\min_{\lambda}\frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}) - \frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}^t) = \min_{\lambda}\frac{1}{ET}\sum_{t=1}^{ET}\boldsymbol{\lambda}^T\mathbf{r}^t - \frac{1}{ET}\sum_{t=1}^{ET}(\boldsymbol{\lambda}^t)^T\mathbf{r}^t \tag{18}$$

$$\leq \frac{B\log(|A|+1)}{\eta_\theta ET} + \eta_\theta\rho^2 B \tag{19}$$

Hence, we conclude that

$$\min_{\boldsymbol{\lambda}}\frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}) - \min_w \frac{1}{ET}\sum_{t=1}^{ET} G(w;\boldsymbol{\lambda}^t) \leq \frac{1}{T}\sum_{t=1}^{T}\frac{\kappa}{\gamma+t-1}\left(\frac{2\max_i C_i}{(1+B)\mu} + \frac{(1+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{20}$$

$$+ \frac{B\log(|A|+1)}{\eta_\theta ET} + \eta_\theta\rho^2 B \tag{21}$$

By Jensen's Inequality, $G(\frac{1}{ET}\sum_{t=1}^{ET} w^t; \boldsymbol{\lambda}) \leq \frac{1}{ET}\sum_{t=1}^{ET} G(w^t; \boldsymbol{\lambda})$. Therefore, we have

$$\min_{\boldsymbol{\lambda}} G(\bar{w}; \boldsymbol{\lambda}) - \min_{w} G(w; \bar{\boldsymbol{\lambda}}) \leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma + t - 1}\left(\frac{2\max_i C_i}{(1+B)\mu} + \frac{(1+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{22}$$

$$+ \frac{B\log(|A|+1)}{\eta_\theta ET} + \eta_\theta \rho^2 B \tag{23}$$

Let $C_1 = \max_i C_i$ and $C_2 = \max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]$, we get Theorem 1. □

Proof for corollary 1. Note that $\log(t+1) \leq \sum_{n=1}^{t} \frac{1}{n} \leq \log(t) + 1$ Let

$$\mathcal{C} = \frac{2\max_i C_i}{(1+B)\mu} + \frac{(1+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2] \tag{24}$$

we have

$$\min_{\boldsymbol{\lambda}} G(\bar{w}; \boldsymbol{\lambda}) - \min_{w} G(w; \bar{\boldsymbol{\lambda}}) \leq \frac{\kappa\mathcal{C}}{T}\left(\log(\gamma + T - 1) + 1 - \log(\gamma + 1)\right) + \frac{B\log(|A|+1)}{\eta_\theta ET} + \eta_\theta\rho^2 B \tag{25}$$

Denote the right hand side as ν_T. Pick $\eta_\theta = \frac{\nu}{2\rho^2 B}$ and $T \geq \frac{1}{\nu(\gamma+1) - 2\kappa\mathcal{C}}\left(\frac{4\rho^2 B^2 \log(|A|+1)(\gamma+1)}{\nu E} + 2\kappa\mathcal{C}(\gamma - 1)\right)$.

$$\nu_T \leq \frac{\kappa\mathcal{C}}{T}\frac{\gamma + T - 1}{\gamma + 1} + \frac{2\rho^2 B^2 \log(|A|+1)}{\nu ET} + \frac{\nu}{2} \tag{26}$$

$$= \frac{1}{T}\frac{\kappa\mathcal{C}(\gamma-1)\nu E + 2\rho^2 B^2 \log(|A|+1)(\gamma+1)}{\nu E(\gamma+1)} + \frac{\kappa\mathcal{C}}{\gamma+1} + \frac{\nu}{2} \tag{27}$$

$$\leq \frac{\nu E(\nu(\gamma+1) - 2\kappa\mathcal{C})}{4\rho^2 B^2 \log(|A|+1)(\gamma+1) + 2\kappa\mathcal{C}(\gamma-1)\nu E}\frac{\kappa\mathcal{C}(\gamma-1)\nu E + 2\rho^2 B^2 \log(|A|+1)(\gamma+1)}{\nu E(\gamma+1)} + \frac{\kappa\mathcal{C}}{\gamma+1} + \frac{\nu}{2} \tag{28}$$

$$= \frac{\nu E(\nu(\gamma+1) - 2\kappa\mathcal{C})}{2\nu E(\gamma+1)} + \frac{\kappa\mathcal{C}}{\gamma+1} + \frac{\nu}{2} \tag{29}$$

$$= \frac{\nu}{2} + \frac{\nu}{2} \tag{30}$$

$$= \nu \tag{31}$$

□

E Proof for Theorem 3

We first introduce a few lemmas necessary for proof for Theorem 3.

Lemma 2. *Let*

$$\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_{S_k \sim \mathcal{D}_k^{m_k}, \sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k}\sigma_{k,i} l\left(h(x_{k,i}), y_{k,i}\right)\right]$$

then for any $h \in \mathcal{H}$, with probability $1-\delta$, we have

$$|\mathcal{F}(h) - F(h)| \leq 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)} \tag{32}$$

Proof for lemma 2. Lemma 2 directly follows proof for Theorem 2 in Mohri et al. (2019) with $\lambda_k = \frac{1}{K}$.

□

Lemma 3. *Let*

$$\mathfrak{R}_a(\mathcal{H}) = \mathbb{E}_{S_k \sim \mathcal{D}_k^{m_k}, \sigma}\left[\sup_{h\in\mathcal{H}} \sum_{k=1}^{K} \frac{1}{m_a} \sum_{a_i=a} \sigma_{k,i} l\left(h(x_{k,i}), y_{k,i}\right)\right]$$

then for any $h \in \mathcal{H}$ and $a \in A$, with probability $1-\delta$, we have

$$|\mathfrak{r}_a(h) - \mathbf{r}_a(h)| \leq 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_a}\sqrt{\frac{K}{2}\log(1/\delta)} \tag{33}$$

Lemma 4 (Lemma 1 in Agarwal et al. (2018))**.** *Let $(\bar{w}, \bar{\boldsymbol{\lambda}})$ is a ν-approximate saddle point, then*

$$\bar{\boldsymbol{\lambda}}^T \mathbf{r}(\bar{w}) \geq B \max_{a\in A} \mathbf{r}_a(\bar{w})_+ - \nu \tag{34}$$

where $x_+ = \max\{x, 0\}$.

Lemma 5 (Lemma 2 in Agarwal et al. (2018))**.** *For any w such that $\mathbf{r}(w) \leq \mathbf{0}_{|A|}$, $F(\bar{w}) \leq F(w) + 2\nu$.*

Lemma 6. *Assume there exists w^* satisfies $\mathbf{r}(w^*) \leq \mathbf{0}_{|A|}$, we have*

$$B \max_{a\in A} \mathbf{r}_a(\bar{w})_+ \leq M + 2\nu \tag{35}$$

Proof for lemma 6. Note that

$$\begin{aligned}
F(\bar{w}) + B\max_{a\in A} \mathbf{r}_a(\bar{w})_+ - \nu &\leq F(\bar{w}) + \bar{\boldsymbol{\lambda}}^T \mathbf{r}(\bar{w}) && (36)\\
&= G(\bar{w}, \bar{\boldsymbol{\lambda}}) && (37)\\
&\leq \min_{w} G(w, \bar{\boldsymbol{\lambda}}) + \nu && (38)\\
&\leq G(w^*, \bar{\boldsymbol{\lambda}}) + \nu && (39)\\
&= F(w^*) + \bar{\boldsymbol{\lambda}}^T \mathbf{r}(w^*) + \nu && (40)\\
&\leq F(w^*) + \nu. && (41)
\end{aligned}$$

Therefore, we have

$$F(\bar{w}) \leq F(w^*) + 2\nu. \tag{42}$$

Hence,

$$\begin{aligned}
B\max_{a\in A} \mathbf{r}_a(\bar{w})_+ &\leq F(w^*) - F(\bar{w}) + 2\nu && (43)\\
&\leq M + 2\nu, && (44)
\end{aligned}$$

□

Note that Lemma 6 tells us when there exists a solution for problem 2, the empirical fairness constraint violates by at most an error of $\frac{M+2\nu}{B}$. In other words, this guarantees that our algorithm 1 always output a model when problem 2 has a solution.

Now we provide proof for Theorem 3.

Proof for Theorem 3. When there exists a solution to problem 2: w^*, by lemma 2, 6, we have

$$\begin{aligned}
\mathcal{F}(\bar{w}) &\leq F(\bar{w}) + 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)} && (45)\\
&\leq F(w^*) + 2\nu + 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)} && (46)\\
&\leq \mathcal{F}(w^*) + 2\nu + 4\mathfrak{R}_m(\mathcal{H}) + \frac{2M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)}. && (47)
\end{aligned}$$

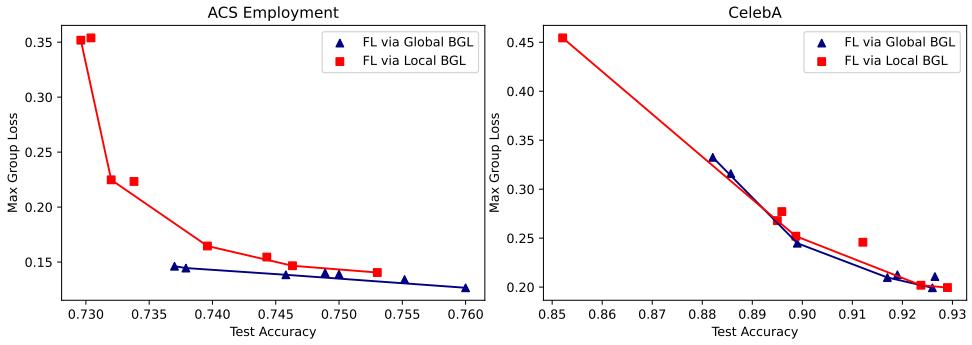

Figure 2:

Combined with lemma 3, 6, we have

$$\mathfrak{r}_a(\bar{w}) \leq \mathbf{r}_a(\bar{w}) + 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_a}\sqrt{\frac{K}{2}\log(1/\delta)} \tag{48}$$

$$\leq \frac{M+2\nu}{B} + 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_a}\sqrt{\frac{K}{2}\log(1/\delta)} \tag{49}$$

Therefore, Theorem 3 holds in this case.

When there doesn't exist a solution to problem 2, Algorithm 1 outputs $\bar{w}$ only when $\max_{a\in A} \mathbf{r}_a(\bar{w}) \leq \frac{M+2\nu}{B}$. In certain scenario, we are still able to obtain

$$\mathfrak{r}_a(\bar{w}) \leq \frac{M+2\nu}{B} + 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_a}\sqrt{\frac{K}{2}\log(1/\delta)} \tag{50}$$

by applying lemma 3. Since w^* doesn't exist,

$$\mathcal{F}(\bar{w}) \leq \mathcal{F}(w^*) + 2\nu + 4\mathfrak{R}_m(\mathcal{H}) + \frac{2M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)} \tag{51}$$

holds vacuously.

Therefore, Theorem 3 holds for both cases.

□

F LOCAL VS GLOBAL FAIRNESS CONSTRAINT

In federated learning, previous work has shown it is not feasible to use local fairness metrics to approximate global fairness metrics. In other words, applying fair training locally at each data silo and aggregate the resulting model is not able to provide strong fairness guarantee at the global level with the same fairness definition (Zeng et al., 2021). In this section, we present and compare the relationship between test accuracy and max group loss under local BGL constraint and global BGL constraint. The results are shown in Figure 2 for both datasets. On ACS Employment dataset, compared to the proposed method, FL via local BGL achieves higher maximum group loss given the same accuracy. Contrary to what is shown in Zeng et al. (2021), on both datasets, even with local BGL constraint, fairness aware federated learning with proper hyper parameters yields a more fair and accurate model than FedAvg. The gap between fairness guarantee when optimizing global and local BGL constraint potentially depends on the data heterogeneity level across the silos. In future work, we are interested in investigating the relation between data heterogeneity and differences in the local vs. global BGL constraint.