2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) | 979-8-3503-4950-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/SaTML59370.2024.00015

Fair Federated Learning via Bounded Group Loss

Shengyuan Hu
Carnegie Mellon University
shengyuanhu@cmu.edu

Zhiwei Steven Wu
Carnegie Mellon University
zstevenwu@cmu.edu

Virginia Smith
Carnegie Mellon University
vsmith@cmu.edu

***Abstract*—Fair prediction across protected groups is an important consideration in federated learning applications. In this work we propose a general framework for provably fair federated learning. In particular, we explore and extend the notion of Bounded Group Loss as a theoretically-grounded approach for group fairness that offers favorable trade-offs between fairness and utility relative to prior work. Using this setup, we propose a scalable federated optimization method that optimizes the empirical risk under a number of group fairness constraints. We provide convergence guarantees for the method as well as fairness guarantees for the resulting solution. Empirically, we evaluate our method across common benchmarks from fair ML and federated learning, showing that it can provide both fairer and more accurate predictions than existing approaches in fair federated learning.**

***Index Terms*—Fairness, Federated Learning**

I. 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 [1, 2, 3]. Many real world FL applications require performing fair prediction across protected groups (e.g., age, gender, race) where the data of each group is distributed across different silos. For example, in applications of cross-silo FL such as learning across hospitals or financial institutions, it is natural to consider fairness constraints with respect to subgroups of patients or users [4, 5, 6]. While many methods have been proposed to incorporate group fairness constraints in centralized settings [e.g., 7, 8, 9, 10], group fairness in cross-silo federated settings remains relatively unexplored.

Unfortunately, existing approaches that have been proposed for group fair FL rely on solving objectives that *equalize* the losses across groups [4, 6, 11, 12]. Such approaches have a number of deficiencies. First, as prediction difficulty can vary between groups, these methods may artificially cause one group's utility to drop in order to enforce equal prediction quality between two groups. For example, as we show in Figure 1, when there is noise in the data from one group, enforcing equal loss may increase the loss of the other group, even if the data of the other group does not change—resulting in a model with low utility. Prior works in centralized fair ML [13, 14] have similarly shown that fairness notions like Demographic Parity and Equalized Odds [7] can harm the utility of both groups. In practice, FL applications often have strict utility constraints, and significantly compromising the performance of one or multiple groups may not be desired. For example, when the silos are hospitals that own private patient data, it may be critical to ensure that the worst group loss according to some protected demographic (e.g., age, race, gender) is no worse than a certain threshold. Finally, despite promising empirical performance in certain settings, prior works in fair FL typically lack formal guarantees surrounding the resulting fairness of the solutions (Section II). This is problematic as it is unclear how the methods may perform in real-world FL deployments; as we show, in practice existing heuristics for fair FL can in fact result in solutions that not only sacrifice utility but also fail to provide the fair performance they seek to optimize (Section V).

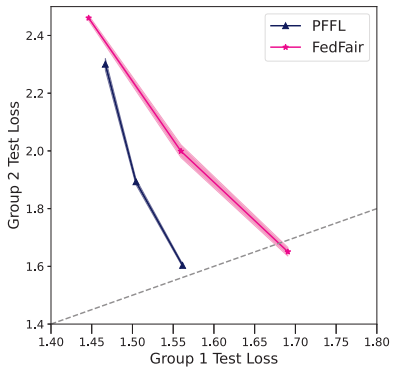

Fig. 1: Existing fair FL approaches that equalize losses may significantly decrease utility. Here we fit a linear regression model with synthetic data from two groups. While using a solver that equalizes two group's losses (FedFair [4]) reduces the loss gap, it results in strictly worse utility for both groups and roughly the same gap as our method (PFFL).

In this work, we instead propose *Bounded Group Loss (BGL)* [15] and its variations as compelling fairness criteria for federated learning. Instead of enforcing equal prediction quality between two groups, BGL sets an upper bound for the loss associated with every protected group, thus ensuring that the worst group's performance meets a pre-defined threshold. As a result, BGL can prevent the model performance from drastically dropping in order to satisfy fairness criteria. Beyond empirical benefits, BGL also has strong theoretical guarantees; the scalable method we propose (PPFL) provably finds the optimal predictor in a hypothesis class subject to the criterion of BGL. Empirically, we demonstrate the effectiveness of our approach relative to existing methods on common benchmarks

979-8-3503-4950-4/24/$31.00 ©2024 IEEE
DOI 10.1109/SaTML59370.2024.00015

from both fair machine learning and federated learning. We summarize our main contributions below:

- We propose a novel *group fair cross-silo federated learning* framework for a range of group fairness notions. Our framework models the fair FL problem as a saddle point optimization problem and leverages variations of Bounded Group Loss [15] to capture common forms of group fairness. We also extend BGL to consider a new fairness notion called Conditional Bounded Group Loss (CBGL), which may be of independent interest and utility in non-federated settings.
- We propose a scalable federated optimization method for our group fair FL framework. We provide a regret bound analysis for our method under convex machine learning objectives to demonstrate formal convergence guarantees. Further, we provide fairness and generalization guarantees on the model for a variety of fairness notions.
- Finally, we evaluate our method on common benchmarks used in fair machine learning and federated learning. In all settings, we find that our method can improve fairness in terms of worst group performance without drastically compromising the overall utility. Additionally, though we do not directly optimize certain group fairness constraints such as Demographic Parity and Equal Opportunity, we find that our method provides competitive fairness-utility trade-offs relative to existing approaches evaluated on these metrics, including those proposed specifically for these criteria.

The remainder of the paper is organized as follows. We discuss related work in fair machine learning and federated learning in Section II. In Section III, we formalize our fairness definition via *Bounded Group Loss* and provide intuition for the use of the objective for addressing group fair FL. In Section IV we present a scalable algorithm to solve the proposed objective in federated settings, and provide formal convergence and fairness guarantees for our objective and algorithm. In Section V we evaluate our approach on benchmarks from fair and federated learning, demonstrating that our method provides favorable fairness-utility trade-offs in practice across a number of fairness metrics relative to existing approaches.

II. Background and Related Work

Algorithmic fairness in machine learning aims to identify and correct bias in the learning process. In federated learning, definitions of fairness can take many forms. A common notion of fairness is representation parity [16], whose application in FL typically requires the model's performance across all clients to have small variance [17, 18, 19, 20, 21, 22, 23]. In this work we instead focus on notions of *group fairness*, in which every data point in the federated network belongs to some (possibly) protected group, and we aim to find a model that doesn't introduce bias towards any group. As shown in Figure 2, in cross-silo federated settings where data is distributed across different data silos such as hospitals or financial institutions, applying fair methods locally only ensures fairness for each silo rather than the entire population. Developing effective and efficient techniques for globally group fair FL is thus an important area of study.

To develop principled approaches for group fair FL, a natural starting point would be to leverage existing work in the centralized setting. However, in order to find the optimal predictor subject to a fairness constraint such as equalized odds [7], many centralized algorithms require solvers for optimal cost-sensitive classifiers [9], the Bayes-optimal predictor [7], or a multi-calibrated predictor [24]. These solvers either do not exist or are too demanding to assume in a federated setting. In addition, achieving fairness in federated settings inherits common challenges that already exist in centralized settings. For example, the criterion of loss parity, which requires equality of losses across groups, often leads to non-convexity even if the underlying loss function is convex [25].

In response to these issues, recent works have proposed various heuristics for achieving group fairness in FL. Zeng et al. [12] consider a bi-level optimization objective that minimizes the difference between each group's loss while finding an optimal global model. Similarly, several works propose using a constrained optimization problem that aims to find the best model subject to an upper bound on the group loss difference [4, 6, 26, 27]. Unlike these approaches, our method focuses on a fairness constraint based on upperbounding the loss of each group with a constant, which we show can help to improve fairness while avoiding significant drops in utility relative to approaches that aim to equalize performance. More closely related to our work, Papadaki et al. [28] weighs the empirical loss given each group by a trainable vector $\boldsymbol{\lambda}$ and finds the best model for the worst case $\boldsymbol{\lambda}$. Though similar to our method for $\zeta = 0$, this approach fails to achieve both strong utility and fairness performance under non-convex loss functions (see Appendix V-D). Zhang et al. [29] also propose a similar objective to learn a model with unified fairness. Among these works, Zeng et al. [12] and Cui et al. [6] also provide simplified convergence and fairness guarantees for their methods. However, these works lack formal analyses around the convergence for arbitrary convex loss functions as well as the behavior of the fairness constraint over the true data distribution. Ours is the first work we are aware to provide such guarantees in the context of group fair federated learning.

III. Fair FL via Bounded Group Loss

In this section we first formalize the group fair federated learning problem and a fairness-aware objective solving this problem (Section III-A). We then provide several examples of group fairness based on the notion of BGL and show how to incorporate them into our framework (Section III-B).

A. Setup: Group Fair Federated Learning

Following standard federated learning scenarios [1], we consider a network with K different clients. Each client $k \in [K]$ has access to training data $\hat{\mathcal{D}}_k := \{(x_i, y_i, a_i)\}_{i=1,\cdots,m_k}$

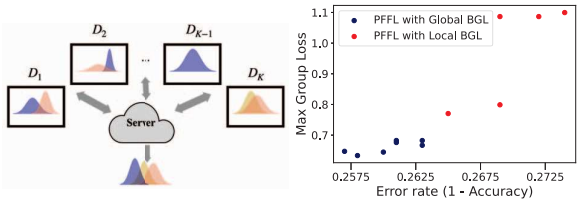

Fig. 2: *Left:* Due to data heterogeneity in federated networks, the data distribution conditioned on each protected attribute (specified by different colors) may differ across clients. The purpose of fair federated learning is to learn a model that provides fair prediction on the entire true data distribution. *Right:* Empirical results on ACS dataset also show that training with local fairness constraint induce both higher error rate and worse fairness guarantee then training with global fairness constraint. This motivates the development of methods that can enable global group fairness in federated settings.

sampled from the true data distribution $\mathcal{D}_k$, where x_i is an observation, $y_i \in Y$ is the label, $a_i \in A$ is the protected attribute. Let the hypothesis class be $\mathcal{H}$ and for any model $h \in \mathcal{H}$, and define the loss function on data (x, y, a) to be $l(h(x), y)$. Federated learning applications typically aim to solve:

$$\min_{h\in\mathcal{H}} \mathcal{F}(h) = \min_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[l(h(x), y)\right] . \tag{1}$$

In practice, $\mathcal{D}_k$ is estimated by observing $\{(x_i, y_i, a_i)\}_{i=1,\cdots,m_k}$, and we solve the empirical risk:

$$\min_{h\in\mathcal{H}} F(h) = \min_{h\in\mathcal{H}} \frac{1}{K}\sum_{k=1}^{K} \frac{1}{m_k}\sum_{i=1}^{m_k} l(h(x_{k,i}), y_{k,i}) . \tag{2}$$

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 $f_k(h)$ intermittently in the remainder of the paper.

Fairness via Constrained Optimization. When a centralized dataset is available, a standard approach to learn a model that incorporates fairness criteria is to solve a constrained optimization problem where one tries to find the best model subject to a fairness notion [15, 30]. We formalize a similar learning problem in the federated setting, solving:

$$\min_{h\in\mathcal{H}} F(h) \quad \text{subject to } \mathbf{R}(h) \le \boldsymbol{\zeta}, \tag{3}$$

where $\mathbf{R}(h), \boldsymbol{\zeta} \in \mathbb{R}^Z$ encodes the constraint set on h. For instance, the z-th constraint could be written as $\mathbf{R}_z(h) \le \zeta_z$ where ζ_z is a fixed constant. This formulation is commonly used to satisfy group fairness notions such as equal opportunity, equalized odds [7], and minimax group fairness [14].

To solve the constrained optimization problem (3), a common method is to use Lagrange multipliers. In particular, let $\boldsymbol{\lambda} \in \mathbb{R}^Z_+$ be a dual variable with positive values and assume $\boldsymbol{\lambda}$ has $\|\cdot\|_1$ at most B. The magnitude of B could be viewed as the regularization strength for the constraint term. Objective (3) can then be converted into the following saddle point optimization problem:

$$\min_{w} \max_{\boldsymbol{\lambda}\in\mathbb{R}^Z_+, \sum_i \boldsymbol{\lambda}^Z_{i=1}\le B} G(w;\boldsymbol{\lambda}) = \beta F(w) + \boldsymbol{\lambda}^T \mathbf{r}(w) , \tag{Main Objective}$$

where the q-th index of $\mathbf{r}$ encodes the q-th constraint from $\mathbf{R}$ (i.e., $\mathbf{r}_q(w) := \mathbf{R}_q(w) - \zeta_q$) and β is a fixed constant. In other words, the objective finds the best model under the scenario where the fairness constraint is most violated (i.e., the regularization term is maximized).

There are two steps needed in order to provide meaningful utility and fairness guarantees for the model found by solving Main Objective: (1) showing that it is possible to recover a solution close to the *'optimal'* solution, (2) providing an upper bound for both the risk ($F(w)$) and the fairness constraint ($\mathbf{r}(w)$) given this solution. To formally define what an *'optimal'* solution is, in this work we focus on the case where $G(w;\boldsymbol{\lambda})$ is convex in w for any fixed $\boldsymbol{\lambda}$.

Since G is linear in $\boldsymbol{\lambda}$, 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., $G(w_0;\boldsymbol{\lambda}^*) \ge G(w_0;\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$. When G is convex in w, given a fixed $\boldsymbol{\lambda}_0$, there exists w^* that satisfies $w^* = \arg\min_w G(w;\boldsymbol{\lambda}_0)$, i.e., $G(w^*;\boldsymbol{\lambda}_0) \le G(w;\boldsymbol{\lambda}_0)$ for all w. Therefore, $(w^*, \boldsymbol{\lambda}^*)$ is a saddle point of $G(\cdot;\cdot)$, which is denoted as the optimal solution in our setting.

B. Formulation: Bounded Group Loss and Variants

Many prior works in fair FL consider instantiating $\mathbf{R}(h)$ in (3) as a constraint that bounds the difference between any two groups' losses, which is a common technique used to enforce group fairness notions such as equalized odds and demographic parity [4, 6, 12]. This results in $G(w;\boldsymbol{\lambda})$ becoming nonconvex in w [25]. Such nonconvexity can be problematic as it increases the likelihood that a solver will find a local minima that either does not satisfy the fairness constraint or achieves poor utility. Instead of enforcing equity between the prediction quality of any two groups, we explore using a constraint based on *Bounded Group Loss (BGL)* [15] which enforces an upper bound for all groups's losses, and propose new variants that can retain convexity assumptions while satisfying meaningful fairness notions. We explore three instantiations of group fairness constraints $\mathbf{R}(h)$ below.

Instantiation 1 (Bounded Group Loss). We begin by considering fairness via the Bounded Group Loss (defined below), which was originally proposed by Agarwal et al. [15]. Different from applying Bounded Group Loss in a centralized setting, BGL in the context of federated learning requires that for any group $a \in A$, the average loss for *all* data belonging to group a is below a certain threshold. As we discuss in Section IV this (along with general constraints of FL such as communication) necessitates the development of novel solvers and analyses for the objective.

Definition III.1 (Bounded Group Loss (BGL) Agarwal et al. [15])**.** A classifier h satisfies Bounded Group Loss (BGL) at level ζ under distribution $\mathcal{D}$ if for all $a \in A$, we have $\mathbb{E}\left[l(h(x), y)|A = a\right] \leq \zeta$.

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 $\mathbf{r}_a(h) := \sum_{k=1}^{K}\mathbf{r}_{a,k}(h) \leq 0$, where

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

Benefits of BGL. BGL enforces an upper bound for all groups' losses, and thus ensures the prediction quality for each group to meet a pre-specified level. Compared to other common fairness criteria such as equalized odds and loss parity, BGL has two main advantages. First, BGL ensures convexity in the problem objective, which facilitates provable guarantees in the federated setting. Further, when the data are not equally predictive across groups, BGL can avoid artificial drops in accuracy across every group for the purpose of matching performance of the worst group (see Section V).

Instantiation 2 (Conditional Bounded Group Loss). In some applications one needs a stronger fairness notion beyond ensuring that no group's loss is too large. For example, in the scenario of binary classification, a commonly used fairness requirement is equalized true positive rate or false positive rate [7]. In the context of optimization for arbitrary loss functions, a natural substitute is equalized true / false positive loss. In other words, any group's loss conditioned on positively / negatively labeled data should be equalized. Therefore, similar to BGL, we propose a novel fairness definition known as *Conditional Bounded Group Loss (CBGL)* defined below:

Definition III.2 (Conditional Bounded Group Loss (CBGL))**.** A classifier h satisfies Conditional Bounded Group Loss (CBGL) for $y \in Y$ at level ζ_y under distribution $\mathcal{D}$ if for all $a \in A$, $\mathbb{E}\left[l(h(x), y)|A = a, Y = y\right] \leq \zeta_y$.

In practice, similar to Equation 4, we could define the empirical CBGL by viewing the tuple (a, y) as a group and take the difference between the average of all examples belonging to one group and its pre-defined threshold. Note that satisfying CBGL for all Y is a strictly harder problem than satisfying BGL alone. In fact, we can show that a classifier that satisfies CBGL at level $[\zeta_y]_{y\in Y}$ also satisfies BGL at level $\mathbb{E}_{y\sim\rho_a}[\zeta_y]$ where ρ_a be the probability density of labels for all data from group a.

Relationship between CBGL and Equalized Odds. For binary classification tasks in centralized settings, a common fairness notion is Equalized Odds (EO) [7], which requires the True/False Positive Rate to be equal for all groups. Our CBGL definition can be viewed as a relaxation of EO. Consider a binary classification example where $Y = \{0, 1\}$. Let the loss function l be the 0-1 loss. CBGL requires classifier h to satisfy $\Pr[h(x) = y|Y = y_0, A = a] \leq \zeta_{y_0}$ for all $a \in A$ and $y_0 \in Y$. EO requires $\Pr[h(x) = y|Y = y_0, A = a]$ to be the same for all $a \in A$ given a fixed y_0, which may not be feasible if the hypothesis class $\mathcal{H}$ is not rich enough. Instead of requiring equity of each group's TPR/FPR, CBGL only imposes an upper bound for each group's TPR/FPR. Similar to the comparison between BGL and loss parity, CBGL offers more flexibility than EO since it does not force an artificial increase on FPR or FNR when a prediction task on one of the protected groups is much harder. In addition, for applications where logistic regression or DNNs are used (e.g., CV, NLP), it is uncommon to use the 0-1 loss in the objective. Thus, CBGL can provide a relaxed notion of fairness for more general loss functions whose level of fairness can be flexibly tuned.

Instantiation 3 (MinMax Fairness). Recently, Papadaki et al. [28] proposed a framework called FedMinMax by solving an agnostic fair federated learning framework where the weight is applied to empirical risk conditioned on each group. Note that using BGL as the fairness constraint, our framework could reduce to FedMinMax as a special case by setting $\beta = 0, B = 1$ and $\zeta = 0$.

Definition III.3. Use the same definition of $\mathbf{r}_a(h)$ as we had in Instantiation 1. FedMinMax [28] aims to solve for the following objective:

$$\min_h \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|A|}, \|\boldsymbol{\lambda}\|_1=1} \sum_{a\in A} \boldsymbol{\lambda}_a \mathbf{r}_a(h) \tag{5}$$

Note that a key property of FedMinMax is the constant ζ used to upper bound the per group loss is set to 0. From a constrained optimization view, the only feasible solution that satisfies all fairness constraints for this problem is a model with perfect utility performance since requiring all losses to be smaller than 0 is equivalent to having all of them to be exactly 0. Such a property limits the ability to provide fairness guarantees for FedMinMax. Fixing B and ζ also limits its empirical performance on the relation between fairness and utility, as we will show later in Section V-D.

IV. Provably Fair Federated Learning

In this section, we first propose *Provably Fair Federated Learning (PFFL)*, a simple, scalable solver for our Main Objective, presented in Algorithm 1. We provide formal convergence guarantees for the method in Section IV-B. Given the solution found by PFFL, in Section IV-C we then demonstrate the fairness guarantee for different examples of fairness notions defined in Section III (BGL, CBGL).

A. Algorithm

To find a saddle point for Main Objective, we follow the scheme from Freund and Schapire [31] and summarize our solver for fair FL in Algorithm 1. Our algorithm is based off of FedAvg [1], a common scalable approach in federated learning.

While [15] also follows a similar recipe to ensure BGL, our method needs to overcome additional challenges in the

Algorithm 1 PFFL: Provably Fair Federated Learning

Input: number of each FedAvg rounds T, number of gradient ascent rounds E, SGD lr η_w, exponentiated gradient ascent lr η_θ, model initialization (w_0, $\bar{w} = \mathbf{0}$, $\theta^0 = \mathbf{0}$), slacks in fairness constraints ζ, convergence threshold ν, bound 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 a set of clients S_t.
    for all k in S_t in parallel **do**
      Each client updates weight w_k for J iterations
      $$w_k^{t+1} = w^t - \eta_w \left(\nabla_{w^t} \left(f_k(w^t) + \boldsymbol{\lambda}^T \mathbf{r}(w^t) \right) \right)$$
      Each client sends $g_k^{t+1} = w_k^{t+1} - w_k^t$ and $\mathbf{r}_{a,k}^t$ back to the server.
    end for
    Server computes $w^{t+1} = w^t + \frac{1}{K} \sum_{k=1}^K g_k^{t+1}$.
    Update $\bar{w} = \sum_{t=1}^T w^t$ and set $w^0 = w^T$
  end for
  Server updates θ which would be later used to update the dual variable $\boldsymbol{\lambda}$
  $$\theta_a^{(i+1)} = \theta_a^i + \eta_\theta \sum_k \mathbf{r}_{a,k}^t$$
end for
Server updates $\bar{w} \leftarrow \frac{1}{ET} \bar{w}$
Output $\bar{w}$ if $\max_a \mathbf{r}_a \leq \frac{M+2\nu}{B}$, and null otherwise.

federated settings. In particular, the method in Agarwal et al. [15] optimizes w by performing exact best response, which is in general infeasible when data is distributed across private silos. Our method overcomes this challenge by applying a gradient-descent-ascent style optimization process that utilizes the output of a FL algorithm as an approximation for the best response. In Algorithm 1, we provide an example in which the first step is achieved by using FedAvg to solve $\min_w F(w) + \boldsymbol{\lambda}^T \mathbf{r}(w)$ (L 4-12). After we obtain a global model from a federated training round, we use exponentiated gradient descent to update $\boldsymbol{\lambda}$, following Alg 2 in [15]. This completes one training round. At the end of training, we calculate and return the average iterate as the fair global model.

Note that the ultimate goal in solving Main Objective is to find a w such that it minimizes the empirical risk subject to $\mathbf{r}(w) \leq 0$. Thus, 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. We will show in the Lemma IV.8 that this is always true when there exists a solution w^* for Problem 3. However, it is also worth noting that the Problem 3 is not always feasible. For example when we set $\zeta = 0$, requiring $\mathbf{r}(w) \leq 0$ 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 $\frac{M+2\nu}{B}$.

Privacy Aspects of PFFL. Compared to FedAvg, we note that our solver communicates losses conditioned on each group in addition to model updates. This is common in prior works that solve a min-max optimization problem in FL [e.g., 12, 32]. Although not the main focus of this work, we note that our method could be easily extended to satisfy example-level DP for FL by performing DP-SGD locally for each client (see Appendix C), and can similarly yield natural client-level DP and LDP variants.

Using a different solver. While FedAvg is a natural solver for our Main Objective, prior works have proposed general federated saddle point optimization solvers which could potentially be used [32, 33]. These works either assume strong concavity in the dual variable [33], which does not hold for our objective, or do not support partial participation [32]. These works also require more hyperparameter tuning than our simple framework, which we find to be sufficient to achieve competitive performance in our experiments (Section V). However, for completeness we provide a comparison between our method and Hou et al. [32] in the case of full client participation in Appendix F.

B. Convergence guarantee

Different from Agarwal et al. [15], while our algorithm handles arbitrary convex losses in federated setting by replacing the best response with the FedAvg output, we aim to understand how close our solution is to the actual best response after running finitely many rounds. In this section, we provide a no regret bound style analysis for our PFFL algorithm. To formally measure the distance between the solution found by our algorithm and the optimal solution, we introduce ν-approximate saddle point as a generalization of saddle point (See Remark in Section III-A) defined below:

Definition IV.1. $(\widehat{w}, \widehat{\boldsymbol{\lambda}})$ is a ν-approximate saddle point of G if

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

As an example, the optimal solution $(w^*, \boldsymbol{\lambda}^*)$ is a 0-approximate saddle point of G. To show convergence, we first introduce some basic assumptions below:

Assumption IV.2. Let f_k be μ-strongly convex and L-smooth for all $k = 1, \cdots, K$.

Assumption IV.3. 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 IV.4. 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 $k = 1, \cdots, K$.

These are common assumptions used when proving the convergence for FedAvg [e.g., 34]. Now we present our main theorem of convergence:

Theorem IV.5 (Informal Convergence Guarantee)**.** *Let Assumption IV.2-IV.4 hold. Define* $\gamma=\max\{8\kappa, J\}$, $\kappa=\frac{L}{\mu}$, *step size* $\eta_Q=\frac{2}{(\beta+B)\mu(\gamma+t)}$, $\eta_\theta=\frac{1}{\sqrt{ET}}$, *and assume* $\|\mathbf{r}\|_\infty \leq \rho$. *Letting* $\bar{w}=\frac{1}{ET}\sum_{t=1}^{ET} w^t$, $\bar{\boldsymbol{\lambda}}=\frac{1}{ET}\sum_{t=1}^{ET}\boldsymbol{\lambda}^t$, $\boldsymbol{\Delta_T}=\max_{\boldsymbol{\lambda}} G(\bar{w};\boldsymbol{\lambda})-\min_w G(w;\bar{\boldsymbol{\lambda}})$, *for constant* C *we have:*

$$\boldsymbol{\Delta_T} \leq \frac{1}{T}\sum_{t=1}^{T}\frac{\kappa}{\gamma+t-1}C+\frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}}. \quad (7)$$

The upper bound in Equation 7 consists of two parts: (1) the error for the FedAvg process to obtain $\bar{w}$ which is a term of order $\mathcal{O}(\log T/T)$; (2) the error for the Exponentiated Gradient Ascent process to obtain $\bar{\boldsymbol{\lambda}}$ which converges with a rate of $\mathcal{O}(1/\sqrt{ET})$. Following Theorem IV.5, we could express the solution of Algorithm 1 as a ν-approximate saddle point of G by picking appropriate η_θ and T:

Corollary IV.6. *Let* $T \geq \frac{2\kappa C(\gamma-1)}{\nu(\gamma+1)-2\kappa C}$ *and* $E \geq \frac{2B^2(\nu(\gamma+1)-2\kappa C)(\log(Z+1)+\rho^2)^2}{\nu^2\kappa C(\gamma-1)}$, *then* $(\bar{w},\bar{\boldsymbol{\lambda}})$ *is a* ν*-approximate saddle point of* G.

We provide detailed proofs for both Theorem IV.5 and Corollary IV.6 in Appendix A. Unlike the setting commonly considered in prior FedAvg analyses [e.g., 34], in our case the outer minimization problem changes as $\boldsymbol{\lambda}$ gets updated. Thus, our analysis necessitates considering a more general scenario where the objective function could change over time.

C. Fairness guarantee

In the previous section, we demonstrated that our Algorithm 1 could converge and find 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. The ultimate goal for our algorithm is to: (1) learn a model that produces fair predictions on training data, and (2) more importantly, produces fair predictions on test data, i.e., data from federated clients not seen during training.

Before presenting the formal fairness and generalization guarantees, we state the following additional assumption, which is a common assumption for showing the generalization guarantee using the Rademacher complexity generalizations bound [17].

Assumption IV.7. Let $\mathcal{F}$ be upper bounded by constant M.

We first show the fairness guarantee on the training data.

Lemma IV.8 (Empirical Fairness Guarantee)**.** *Let Assumption IV.7 holds. Assume there exists* w^* *satisfies* $\mathbf{r}(w^*) \leq \mathbf{0}_Z$, *we have*

$$\max_j \mathbf{r}_j(\bar{w})_+ \leq \frac{M+2\nu}{B}. \quad (8)$$

Lemma IV.8 characterizes the upper bound for the worst fairness constraint evaluated on the training data. Given a fixed ν, one could increase B to obtain a stronger fairness guarantee, i.e., a smaller upper bound. Combining this with Corollary IV.6, it can be seen that when B is large, additional exponentiated gradient ascent rounds are required to achieve stronger fairness.

Next we formalize the fairness guarantee for the entire true data distribution. 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 $\mathcal{D}$. This result is presented below in Theorem IV.9.

Theorem IV.9 (Full Fairness and Generalization Guarantee)**.** *Let* $Err_{\mathcal{F}}(\hat{w})=\mathcal{F}(\bar{w})-\mathcal{F}(w^*)$ *and* $(\hat{w},\hat{\boldsymbol{\lambda}})$ *be* ν*-approximate saddle point of* G. *Then with probability* $1-\delta$, *either there doesn't exist solution for Problem 3 and Algorithm 1 returns null or Algorithm 1 returns* $\hat{w}$ *that satisfies*

$$Err_{\mathcal{F}}(\hat{w}) \leq 2\nu+4\mathfrak{R}_m(\mathcal{H})+\sqrt{\sum_{k=1}^{K}\frac{2M^2}{m_k K^2}\log\left(\frac{2}{\delta}\right)},$$
$$\mathfrak{r}_j(\bar{w}) \leq \frac{M+2\nu}{B}+Gen_{\mathbf{r},j} \quad (9)$$

where w^* *is a solution for Problem 3 and* $Gen_{\mathbf{r}}$ *is the generalization error.*

The first part for Equation 9 characterizes how well our model performs over the true data distribution compared to the optimal solution. As number of clients K increases, we achieve smaller generalization error. The second part for Equation 9 characterizes how well the fairness constraints are satisfied over the true data distribution. Note that the upper bound could be viewed as the sum of empirical fairness violation and a generalization error. Based on our fairness notions defined in Section III-B, we demonstrate what generalization error is under different fairness constraints $\mathbf{r}$.

Proposition 1 ($\mathbf{r}$ encodes BGL at level ζ). There are in total $|A|$ fairness constraints, one for each group. Define the weighted Rademacher complexity for group a as $\mathfrak{R}_a(\mathcal{H})=\mathbb{E}_{S_k,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{k=1}^{K}\frac{1}{m_a}\sum_{a_{k,i}=a}\sigma_{k,i}l\left(h(x_{k,i}),y_{k,i}\right)\right]$. In this scenario, we have:

$$Gen_{\mathbf{r},a}=2\mathfrak{R}_a(\mathcal{H})+\frac{M}{m_a}\sqrt{\frac{K}{2}\log(2|A|/\delta)}.$$

Note that the fairness constraint for group a under true distribution in Equation 9 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 a stronger formal fairness guarantee compared to any group with less data. It is also worth noting that this generalization error grows as the number of clients K grows. Recall that the generalization error becomes smaller when K grows; combing the two results together provides us a tradeoff between fairness notion of BGL and utility over the true data distribution in terms of K.

Proposition 2 ($\mathbf{r}$ encodes CBGL at level $[\zeta_y]_{y\in Y}$). There are in total $|A||Y|$ fairness constraints, one for

each group and label. Define the weighted Rademacher complexity for group a conditioned on y as $\mathfrak{R}_{a,y}(\mathcal{H}) = \mathbb{E}_{S_k,\sigma}\left[\sup_{h\in\mathcal{H}} \sum_{k=1}^{K} \sum_{a_{k,i}=a,y_{k,i}=y} \frac{\sigma_{k,i}}{m_{a,y}} l\left(h(x_{k,i}), y\right)\right]$ where $m_{a,y}$ is the number of all examples from group a with label y. In this scenario, we have:

$$Gen_{\mathbf{r},(a,y)} = 2\mathfrak{R}_{a,y}(\mathcal{H}) + \frac{M}{m_{a,y}}\sqrt{\frac{K}{2}\log(2|A||Y|/\delta)}.$$

Similar to Proposition 1, in order to achieve strong fairness guarantees for any specific constraint on the true data distribution, we need a sufficient number of samples associated with that constraint.

We provide details and a proof for Theorem IV.9 in Appendix B. Different from the analysis performed in Agarwal et al. [15], we analyze the generalization behaviour in a federated setting where we introduce the generalization bound as a function of number of clients K. We then further formally demonstrate the tension between utility and fairness performance evaluated on the true data distribution induced by K, which has not been studied previously to the best of our knowledge.

V. Experiments

We evaluate our approach on common benchmarks from fair ML and FL, including Communities & Crime, a dataset commonly studied in fair ML [15, 35]; the US-wide ACS PUMS data, a recent group fairness benchmark dataset [36]; and CelebA [37], a common federated learning dataset. We compare our method with prior methods that aim to enforce other fairness notions (Demographic Parity and Equal Opportunity) in terms of performance associated with each group in Section V-A, and explore other fairness metrics like DP and EO relative to prior works in Section V-B. We explore the empirical difference between training with our global BGL constraint vs. local BGL constraints in Section V-C, and directly compare to the special case of FedMinMax in Section V-D. Finally we provide detailed ablation studies of the hyperparameters of our method in Section V-E.

Setup. For all experiments, we evaluate performance metrics for each group on test data that belongs to all silos. To reflect the federated setting, we use heterogeneous data partitions to create data silos. Communities & Crime and ACS Employment is naturally partitioned into states in US; CelebA are manually partitioned in a non-IID manner into a collection of data silos. A detailed description of datasets, models, and partition schemes can be found in Table II. For all experiments, we use grid search for hyperparameter tuning; details of hyperparameters can be found in Section V-E. We also perform experiments under the scenario where each silo is treated as a distinct group in Appendix E.

A. BGL improves worst group performance

We first explore how the performance of each group differs as we use different fair FL methods. 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 PFFL we select the hyperparameter pair (B, ζ) that yields the best performance in terms of both groups. For additional hyperparameter details, please see Appendix V-E. We show group 1 performance, group 2 performance, and their difference in Table I. We compare with prior works that aim to promote equal prediction quality between two groups, including FedFair [4], FairFed [11], and FedFB [12][1]. Since FairFed and FedFB rely on binary label information, we do not consider these baselines for Communities & Crime regression task.

On all datasets, while prior methods are able to reduce the gap between the two group's performance, that usually comes with compromising not only the performance of better performing group but also that of the worst performing group. For example, FedFair achieves near perfect fair prediction across two groups in ACS Employment. However, the test accuracy of both groups suffers from significant drop compared to non fair baseline such as FedAvg, and can in fact result in solutions that are less fair in terms of accuracy difference than the simple FedAvg baseline (e.g., on CelebA). Compared to these approaches, our method can substantially improve the worst group performance across all datasets (Group 2 for ACS Employment and CelebA, Group 1 for Communities & Crime). Meanwhile, we observe that in order to boost the worst group performance, our method does not necessarily sacrifice the other group's performance, compared to the FedAvg baseline.

B. BGL/CBGL evaluated on other fairness notions

Another common fairness notion considered in the fair FL literature is to optimize the difference between every two groups' losses (possibly conditioned on the true label) with the aim of achieving Demographic Parity or Equal Opportunity [4, 6, 7, 12]. Formally, consider the case where the protected attribute set $A = \{0, 1\}$. Define $\Delta_{DP} = |\Pr(h(X) = 1|A = 0) - \Pr(h(X) = 1|A = 1)|$, $\Delta_{EO} = |\Pr(h(X) = 1|A, Y = 0, 1) - \Pr(h(X) = 1|A, Y = 1, 1)|$. These works aim to train a model that achieves small Δ_{DP} or small Δ_{EO}, depending on the fairness constraint selected during optimization. As discussed in Section III-B, CBGL can be viewed as a more general definition of Equal Opportunity and Equalized Odds.

In this section, we compare our method with FedFB [12], FairFed [11], FedFair [4], and FCFL [6], all of which aim to optimize Δ_{DP} and Δ_{EO}. We evaluate Δ_{DP} and Δ_{EO} for all approaches on ACS Employment and CelebA, with results shown in Figure 3. We also conduct additional experiments on COMPAS; results are shown in Appendix D. Similar to Figure 4, we show points lying on the pareto frontier for our method.

Although PFFL with BGL and CBGL was not directly designed for these fairness criteria (i.e., it does not directly

[1]FedFB applies when each client has data from all the groups, which does not hold for our CelebA partitioning.

		FedAvg	AFL	FedFair	FairFed	FedFB	**PFFL (ours)**
ACS-E	Group 1 Acc (↑)	.7599 ± .0009	.7193 ± .0019	.7297 ± .0005	.7533 ± .002	.7413 ± .0022	**.7637 ± .0019**
	Group 2 Acc (↑)	.7375 ± .0002	.6733 ± .003	.7261 ± .0025	.7388 ± .0023	.7173 ± .0025	**.7473 ± .0028**
	Acc Difference (↓)	.0234 ± .0002	.046 ± .0015	**.0025 ± .0013**	.0156 ± .0008	.024 ± .0006	.0164 ± .0018
CelebA	Group 1 Acc (↑)	.9429 ± .0002	**.9481 ± .0021**	.9466 ± 0.0008	.9403 ± .0013	-	.9462 ± 0.0011
	Group 2 Acc (↑)	.9394 ± .0004	.9292 ± .0028	.9394 ± .0016	.9352 ± .002	-	**.9424 ± .0018**
	Acc Difference (↓)	**.0035 ± .004**	.0192 ± .0011	.0072 ± .0009	.0051 ± .0015	-	.0049 ± .0009
Crime	Group 1 RMSE (↓)	.2005 ± .0321	.3041 ± .0707	.2897 ± .0707	-	-	**.1918 ± .02**
	Group 2 RMSE (↓)	.1265 ± .02	.1718 ± .0707	.1865 ± .02	-	-	**.1261 ± .0141**
	RMSE Difference (↓)	.074 ± .0326	.1323 ± .0949	.1032 ± .0894	-	-	**.0657 ± .02**

TABLE I: Comparison between PFFL and prior methods in terms of the test performance metric associated with each group. The method that achieves the best performance is **bolded**. Unlike other works which equalize performance at the expense of utility, we see that PFFL can significantly improve the worst group's test performance while maintaining average performance across groups. We also note that prior works that aim only to equalize performance can behave unexpectedly—sometimes worse than simple baselines such as FedAvg.

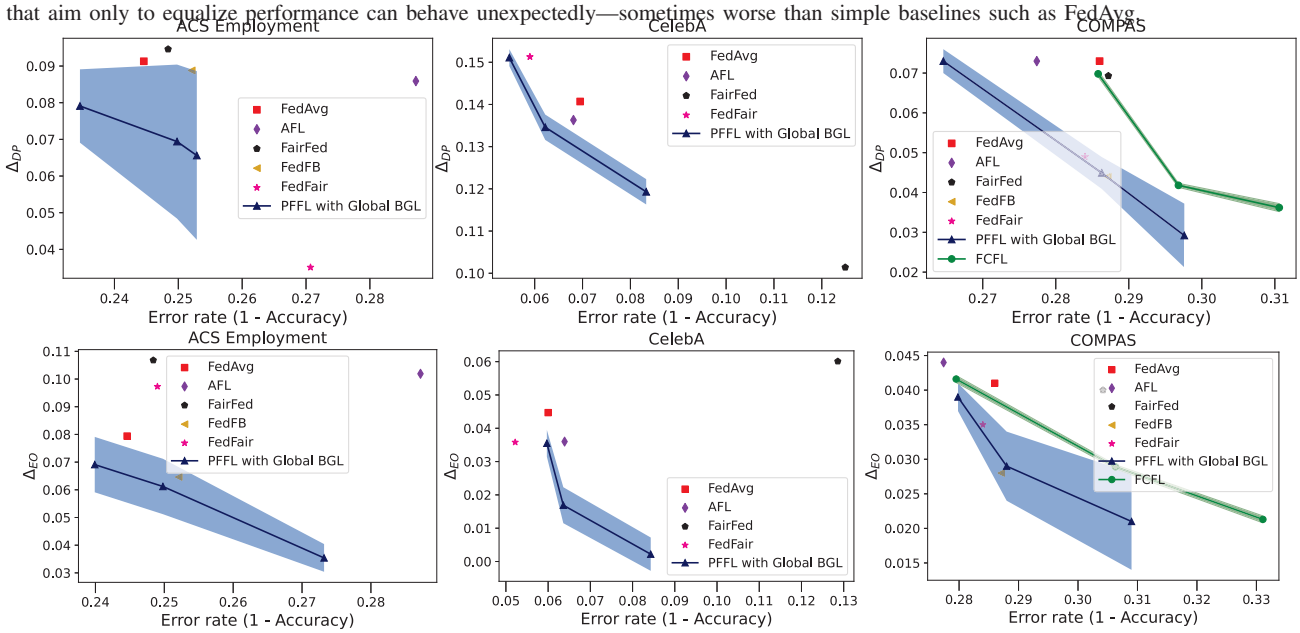

Fig. 3: Comparison between PFFL and prior works on ACS and CelebA for ΔDP and ΔEO. Although PPFL was not directly designed to optimize Demographic Parity/Equal Opportunity, we see that it outperforms the baseline of FedAvg, and performs comparably to/better than prior works designed for these objectives.

enforce the loss or prediction parity of two groups' losses to be close), we see that our method is still able to outperform training a FedAvg baseline, and in fact performs comparably or better than prior methods which were designed specifically for these criteria. Given this empirical performance, studying whether BGL can provide any provable guarantees in terms of Demographic Parity and Equal Opportunity would be an interesting direction of future study.

C. FL with global fairness constraint vs local fairness constraint

In FL, applying fair training locally at each data silo and aggregating the resulting model may not provide strong population-wide fairness guarantees with the same fairness definition [12]. Here, we explore the relationship between test accuracy and max group loss under local BGL and global BGL constraints. The results are shown in Figure 4.

TABLE II: Details of datasets/models used in our experiments.

Dataset	**# of Silos**	**Model**	**Protected Attribute**	**Partition Type**	**Task Type**
ACS	50	Logistic Regression	Race	Natural partition by States	Binary classification
CelebA	40	4-layer CNN	Gender	Manual partition	Binary classification
Communities & Crime	50	Linear Model	Race	Natural partition by States	Linear Regression
COMPAS	10	Logistic Regression	Gender	Manual partition	Binary classification

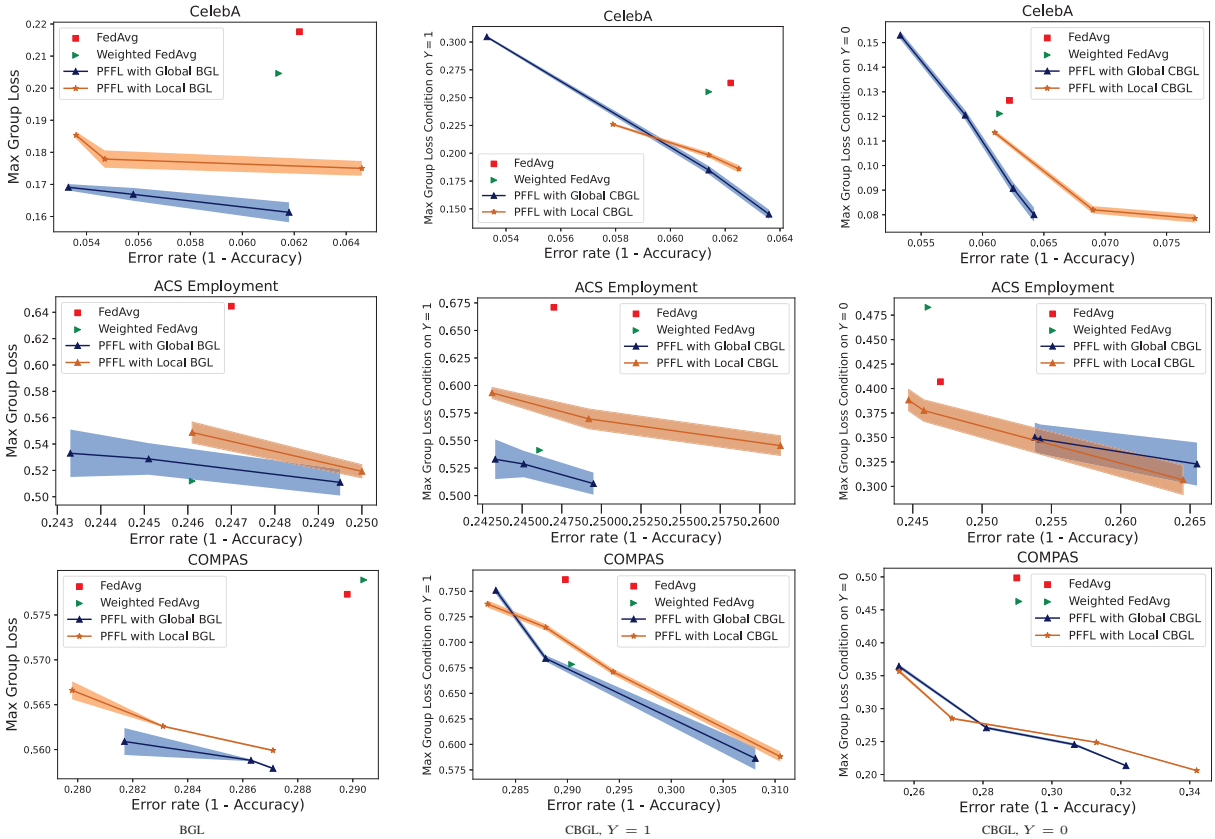

Fig. 4: Empirical results when using BGL (column 1), CBGL for $Y = 1$ (column 2), and CBGL for $Y = 0$ (column 3) on CelebA (row 1), ACS Employment (row 2), and COMPAS (row 3). We find in all settings that our proposed method (PPFL with Global BGL) not only enables a flexible fairness/utility trade-off, but can in fact achieve both stronger fairness and better utility (lower error) than baselines.

On all datasets, we see that there exists a natural tradeoff between error rate and the fairness constraint: when a model achieves stronger fairness (smaller max group loss), the model tends to have worse utility (higher error rate). However, in all scenarios, 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. Meanwhile, for all datasets and fairness metrics, as expected, PFFL with Global BGL achieves improved fairness-utility tradeoffs relative to PFFL with Local BGL. Therefore, our PFFL with Global fairness constraint framework yields a model where utility can coexist with fairness constraints relying on Bounded Group Loss.

D. Comparison with FedMinMax

As mentioned in Section III-B, FedMinMax [28] can be viewed as a special case of our PFFL with BGL with fixed hyperparameters. However, different from our Algorithm 1, the original FedMinMax solver proposes to fix the FedAvg training epoch $T = 1$ whereas our PFFL proposes to training enough number of FedAvg rounds. We compare PFFL with FedMinMax w.r.t worst group accuracy (Table III), max group loss (Figure 6), and demographic parity gap (Figure 5). Similar to Figure 4, we only plot the pareto frontier of our method. Although PFFL with BGL could reduce to FedMinMax, fixing hyperparameters (e.g. β, B, ζ, T) limits FedMinMax's flexibility to trade off between fairness and utility.

		FedMinMax	**PFFL (ours)**
ACS-E	Group 1 Acc (↑)	$.7481 \pm .0052$	$\mathbf{.7637 \pm .0019}$
	Group 2 Acc (↑)	$.7415 \pm .0044$	$\mathbf{.7473 \pm .0018}$
	Acc Difference (↓)	$\mathbf{.0045 \pm 0.0031}$	$.0162 \pm .0018$
Crime	Group 1 RMSE (↓)	$.2175 \pm .0707$	$\mathbf{.1918 \pm .02}$
	Group 2 RMSE (↓)	$.1432 \pm .014$	$\mathbf{.1261 \pm .0141}$
	RMSE Difference (↓)	$.0743 \pm .0689$	$\mathbf{.0657 \pm .02}$

TABLE III: Comparison between PFFL and FedMinMax in terms of test accuracy and RMSE on each group for ACS Employment and Communities & Crime.

E. Hyperparameters

In order to get the fairest model given a certain test error rate, we apply random grid search over two key hyperparameters in our experiment: the strength of regularizer B and the constant used to bound our fairness constraint (ζ when BGL is the fairness constraint and ζ_y when CBGL conditioned on $Y = y$ is the fairness constraint). For all our experiments with respect to PFFL with BGL and PFFL with CBGL, we select $B \in \{0.1, 0.5, 1, 5, 10, 20, 50, 100, 200, 500\}$. For

ACS-E	$B=1$	$B=5$	$B=10$	$B=20$	$B=50$	$B=100$	$B=200$
$\zeta=0.1$	.6974	.6895	.7376	.7226	.7194	.7276	.7255
$\zeta=0.3$	.6785	.7180	.7052	.7408	.7470	.7430	.7425
$\zeta=0.5$	.6917	.7062	.6959	.7120	.7106	.7399	.7348
$\zeta=0.7$	.6774	.7121	.7061	.6995	.6996	.7399	.7216
$\zeta=0.9$	.6961	.6811	.6870	.7197	.7248	.7239	.7058

TABLE IV: Effect of hyperparameters, B and ζ, on the worst group accuracy for ACS Employment. We observe that best worst group accuracy occurs when choosing large B. ζ should be chosen to be close to the actual max group loss.

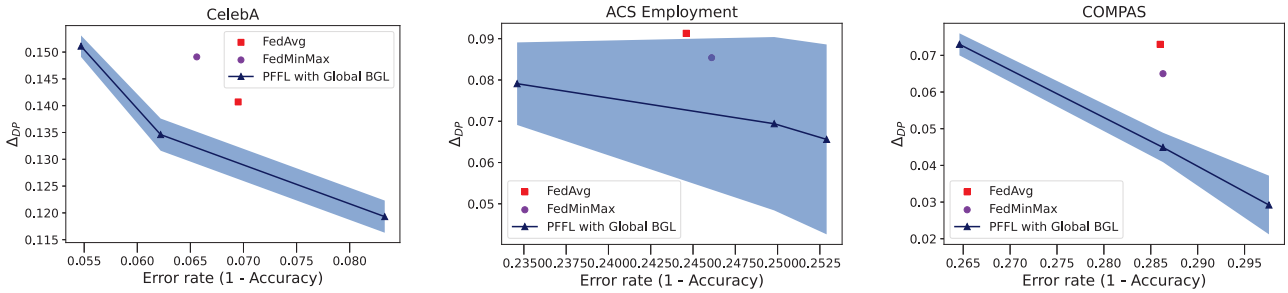

Fig. 5: Comparison between PFFL and FedMinMax in terms of Δ_{DP}. Our method is able to outperform FedMinMax on all three datasets in terms of fairness utility tradeoff.

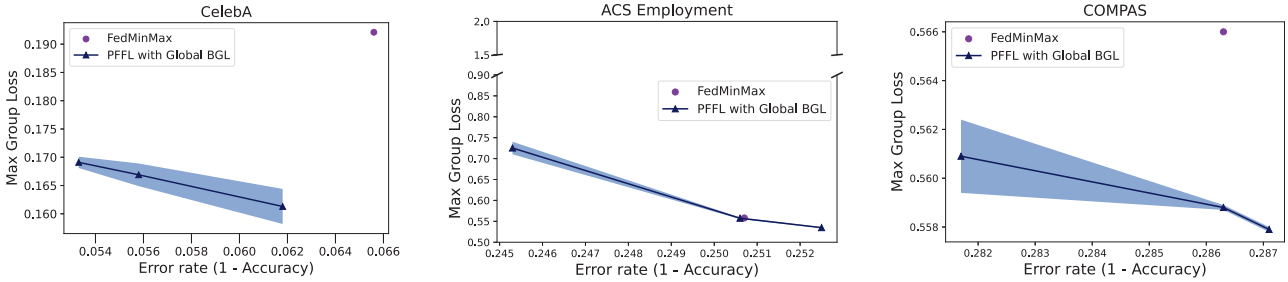

Fig. 6: Comparison with FedMinMax in terms of max group loss. Our method achieves comparable / better performance compared to FedMinMax.

CelebA, we select $\zeta \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$, $\zeta_1 \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$, and $\zeta_0 \in \{0, 0.05, 0.1, 0.15, 0.2\}$. For ACS Employment, we select $\zeta, \zeta_1 \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\zeta_0 \in \{0, 0.1, 0.3, 0.5, 0.7\}$. For Communities & Crime, we select $\zeta \in \{0, 0.01, 0.05, 0.1\}$. For COMPAS, we select $\zeta, \zeta_1 \in \{0, 0.1, 0.3, 0.5, 0.7\}$ and $\zeta_0 \in \{0, 0.1, 0.3, 0.5\}$. We use the same learning rate, total training rounds for all methods including ours and prior works for each dataset. In Table IV, we provide ablation study on ACS-E dataset to study how different hyperparameter combinations affect the worst group performance.

Here we observe two trends: (1) stronger regularization strength (larger B) is beneficial in improving the worst group accuracy for all ζ, and (2) using a ζ too small or too large leads to suboptimal worst group accuracy. In general, while more complex hyperparameter optimization methods could be used, we find that simple grid search over a small set of hyperparameter values with these trends in mind is sufficient to achieve strong empirical performance.

VI. Limitations and Concluding Remarks

In this work, we propose a fair learning objective for federated settings via Bounded Group Loss. We then propose a scalable federated solver to find an approximate saddle point for the objective. Theoretically, we provide convergence and fairness guarantees for our method. Empirically, we show that our method can provide high accuracy and fairness simultaneously across tasks from fair ML and federated learning. In addition to strong empirical performance, ours is the first work we are aware of to provide formal convergence and fairness/generalization guarantees for group fair federated learning with general convex loss functions.

The notion of fairness in machine learning is highly problem-specific, depending not only on aspects of the problem setting but also on values of those who are invoking the fair ML approaches. Indeed, it is well-known that various fairness criteria can in fact be incompatible with one another. Our aim in this work is not to suggest that the notion of Bounded Group Loss and our resulting method for fair federated learning (PFFL) are the only appropriate approaches for ensuring fairness in federated settings, but rather to present the benefits/limitations

of our framework and explain scenarios in which it can be effectively applied in practice. Our results show in particular that fairness can be at odds with other natural concerns in FL, such as overall utility or privacy (Appendix C). While our proposed objective and solver provide improvements relative to existing fair FL approaches along these axes, it is important to not apply our framework (or any other approach for fair ML) blindly, but to carefully consider such trade-offs for the application at hand.

In future work we are interested in investigating additional benefits that could be provided by using our framework, including applications in non-federated settings. We are also interested in extending our analyses to further provide theoretical guarantees for non-convex concave saddle point optimization. Additionally, similar to prior works in group fair FL, our method communicates additional parameters beyond standard non-fair FL (e.g., via FedAvg); while we show that our method is compatible with differentially private training (Appendix C), further studying how privacy interacts with fairness and accuracy in the context of federated learning would be an interesting direction of future work enabled by our framework.

VII. Acknowledgement

This work was supported in part by National Science Foundation Grants IIS2145670 and CCF2107024, a Meta Faculty Award, and the Private AI Collaborative Research Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the National Science Foundation or any other funding agency. ZSW is supported in part by the NSF grant # 1939606.

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 *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[4] L. Chu, L. Wang, Y. Dong, J. Pei, Z. Zhou, and Y. Zhang, "Fedfair: Training fair models in cross-silo federated learning," *arXiv preprint arXiv:2109.05662*, 2021.

[5] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan *et al.*, "Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: Machine learning approach," *JMIR medical informatics*, vol. 9, no. 1, 2021.

[6] S. Cui, W. Pan, J. Liang, C. Zhang, and F. Wang, "Addressing algorithmic disparity and performance inconsistency in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, 2016.

[8] 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*, 2017.

[9] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

[11] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," *arXiv preprint arXiv:2110.00857*, 2021.

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

[13] N. Martinez, M. Bertran, and G. Sapiro, "Minimax pareto fairness: A multi objective perspective," in *International Conference on Machine Learning*, 2020, pp. 6755–6764.

[14] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth, "Minimax group fairness: Algorithms and experiments," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 66–76.

[15] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*. PMLR, 2019.

[16] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*. PMLR, 2018.

[17] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019.

[18] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[19] K. Donahue and J. Kleinberg, "Models of fairness in federated learning," *arXiv preprint arXiv:2112.00818*, 2021.

[20] X. Yue, M. Nouiehed, and R. A. Kontar, "Gifair-fl: An approach for group and individual fairness in federated learning," *arXiv preprint arXiv:2108.02741*, 2021.

[21] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6357–6368.

[22] X. Xu, L. Lyu, X. Ma, C. Miao, C. S. Foo, and B. K. H. Low, "Gradient driven rewards to guarantee fairness in collaborative machine learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 104–16 117, 2021.

[23] W. Chu, C. Xie, B. Wang, L. Li, L. Yin, H. Zhao, and B. Li, "Focus: Fairness via agent-awareness for federated learning on heterogeneous data," *arXiv preprint arXiv:2207.10265*, 2022.

[24] I. Globus-Harris, V. Gupta, C. Jung, M. Kearns, J. Morgenstern, and A. Roth, "Multicalibrated regression for downstream fairness," *CoRR*, vol. abs/2209.07312, 2022.

[25] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," *arXiv preprint arXiv:2012.01696*, 2020.

[26] B. Rodríguez-Gálvez, F. Granqvist, R. van Dalen, and M. Seigel, "Enforcing fairness in private federated learning via the modified method of differential multipliers," *arXiv preprint arXiv:2109.08604*, 2021.

[27] W. Du, D. Xu, X. Wu, and H. Tong, "Fairness-aware agnostic federated learning," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021, pp. 181–189.

[28] A. Papadaki, N. Martinez, M. Bertran, G. Sapiro, and M. Rodrigues, "Minimax demographic group fairness in federated learning," *arXiv preprint arXiv:2201.08304*, 2022.

[29] F. Zhang, K. Kuang, Y. Liu, C. Wu, F. Wu, J. Lu, Y. Shao, and J. Xiao, "Unified group fairness on federated learning," *arXiv preprint arXiv:2111.04986*, 2021.

[30] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning. fairmlbook. org," *URL: http://www. fairmlbook. org*, 2019.
[31] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[32] C. Hou, K. K. Thekumparampil, G. Fanti, and S. Oh, "Efficient algorithms for federated saddle point optimization," *arXiv preprint arXiv:2102.06333*, 2021.
[33] Y. Shen, J. Du, H. Zhao, B. Zhang, Z. Ji, and M. Gao, "Fedmm: Saddle point optimization for federated adversarial domain adaptation," *arXiv preprint arXiv:2110.08477*, 2021.
[34] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.
[35] M. Redmond and A. Baveja, "A data-driven software tool for enabling cooperative information sharing among police departments," *European Journal of Operational Research*, vol. 141, no. 3, pp. 660–678, 2002.
[36] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *Advances in Neural Information Processing Systems*, 2021.
[37] S. Caldas, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings, https://leaf.cmu.edu/," *arXiv preprint arXiv:1812.01097*, 2018.
[38] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.
[39] Z. Liu, S. Hu, Z. S. Wu, and V. Smith, "On privacy and personalization in cross-silo federated learning," *Advances in Neural Information Processing Systems*, 2022.
[40] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
[41] Z. Yang, S. Hu, Y. Lei, K. R. Vashney, S. Lyu, and Y. Ying, "Differentially private sgda for minimax problems," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 2192–2202.
[42] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," *arXiv preprint arXiv:1911.02054*, 2019.

APPENDIX A
PROOF OF THEOREM IV.5

We first present the formal version for Theorem IV.5.

Theorem A.1 (Formal Convergence Guarantee)**.** *Let Assumption IV.2-IV.4 hold. Define* $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, J\}$ *and step size* $\eta_Q = \frac{2}{(\beta+B)\mu(\gamma+t)}$, $\eta_\theta = \frac{1}{\sqrt{ET}}$, *and assume* $\|\mathbf{r}\|_\infty \leq \rho$. *Letting* $\bar{w} = \frac{1}{ET}\sum_{t=1}^{ET} w^t$, $\bar{\boldsymbol{\lambda}} = \frac{1}{ET}\sum_{t=1}^{ET} \boldsymbol{\lambda}^t$, $\boldsymbol{\Delta_T} = \max_{\boldsymbol{\lambda}} G(\bar{w}; \boldsymbol{\lambda}) - \min_w G(w; \bar{\boldsymbol{\lambda}})$, *there exists constant* C_1, C_2 *such that:*

$$\boldsymbol{\Delta_T} \leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1}\left(\frac{2C_1}{(\beta+B)\mu} + \frac{(\beta+B)\mu\gamma}{2}C_2\right) + \frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}}. \tag{10}$$

To prove Theorem A.1, we introduce the following lemma.

Lemma A.2 (Li et al. [34])**.** *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 partial device participation such that* $|S_t| = K$ *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

$$C = \sum_{i=1}^{N} p_i^2\sigma_i^2 + 6L\Gamma + 8(J-1)^2G^2 + \frac{4}{K}J^2G^2.$$

Proof for Theorem A.1. Let $m_{a,k}$ be the number of data with protected attribute a for client k. By Assumption IV.2, we have G_i be $(\beta + \sum_a \boldsymbol{\lambda}_a \frac{m_{a,k}}{m_a})\mu$-strongly convex and $(\beta + \sum_a \boldsymbol{\lambda}_a \frac{m_{a,k}}{m_a})L$-smooth. Since $\|\lambda\|_1 \leq B$, we have G_i be $(\beta+B)\mu$-strongly convex and $(\beta+B)L$-smooth. We first present the regret bound for w^t

$$\frac{1}{ET}\sum_{t=1}^{ET} G(w^t; \lambda^t) - \min_w \frac{1}{ET}\sum_{t=1}^{ET} G(w; \lambda^t) = \frac{1}{ET}\left(\sum_{t=1}^{ET} G(w^t; \lambda^t) - \min_w \sum_{t=1}^{ET} G(w; \lambda^t)\right) \tag{11}$$

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

$$\leq \frac{1}{ET}\left(\sum_{i=0}^{E-1}\left(\sum_{t=1}^{T} G(w^{iT+t}; \lambda^i) - \min_w \sum_{t=1}^{T} G(w; \lambda^i)\right)\right) \tag{13}$$

$$= \frac{1}{E}\sum_{i=0}^{E-1}\left(\frac{1}{T}\sum_{t=1}^{T} G(w^t; \lambda^i) - G^*(\lambda^i)\right) \tag{14}$$

$$\leq \frac{1}{ET}\sum_{i=0}^{E-1}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1}\left(\frac{2C_i}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{15}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1}\left(\frac{2\max_i C_i}{\mu} + \frac{\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{16}$$

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

$$\boldsymbol{\lambda}^T\mathbf{r}^t = \widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{r}}^t \tag{17}$$

for all λ.

By Shalev-Shwartz et al. [38], for any $\widetilde{\boldsymbol{\lambda}}$, we have

$$\sum_{t=1}^{ET} \widetilde{\boldsymbol{\lambda}}^T\widetilde{\mathbf{r}}^t \leq \sum_{t=1}^{ET} (\widetilde{\boldsymbol{\lambda}}^t)^T\widetilde{\mathbf{r}}^t + \frac{B\log(Z+1)}{\eta_\theta} + \eta_\theta\rho^2 BET \tag{18}$$

$$= \sum_{t=1}^{ET} (\boldsymbol{\lambda}^t)^T\mathbf{r}^t + \frac{B\log(Z+1)}{\eta_\theta} + \eta_\theta\rho^2 BET \tag{19}$$

Therefore, we have

$$\min_{\lambda} \frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}) - \frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}^t) = \min_{\lambda} \frac{1}{ET}\sum_{t=1}^{ET} \boldsymbol{\lambda}^T \mathbf{r}^t - \frac{1}{ET}\sum_{t=1}^{ET} (\boldsymbol{\lambda}^t)^T \mathbf{r}^t \tag{20}$$

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

Hence, we conclude that

$$\min_{\boldsymbol{\lambda}} \frac{1}{ET}\sum_{t=1}^{ET} G(w^t;\boldsymbol{\lambda}) - \min_{w} \frac{1}{ET}\sum_{t=1}^{ET} G(w;\boldsymbol{\lambda}^t) \tag{22}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1}\left(\frac{2\max_i C_i}{(\beta+B)\mu} + \frac{(\beta+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) + \frac{B\log(Z+1)}{\eta_\theta ET} + \eta_\theta\rho^2 B \tag{23}$$

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

$$\min_{\boldsymbol{\lambda}} G(\bar{w};\boldsymbol{\lambda}) - \min_{w} G(w;\bar{\boldsymbol{\lambda}}) \leq \frac{1}{T}\sum_{t=1}^{T} \frac{\kappa}{\gamma+t-1}\left(\frac{2\max_i C_i}{(\beta+B)\mu} + \frac{(\beta+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2]\right) \tag{24}$$

$$+ \frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}} \tag{25}$$

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

Next we prove Corollary IV.6.

Proof for corollary IV.6. Note that $\log(t+1) \leq \sum_{n=1}^{t} \frac{1}{n} \leq \log(t)+1$. Let

$$\mathcal{C} = \frac{2\max_i C_i}{(\beta+B)\mu} + \frac{(\beta+B)\mu\gamma}{2}\max_i \mathbb{E}[\|w^{1,i} - w^{*,i}\|^2] \tag{26}$$

we have

$$\min_{\boldsymbol{\lambda}} G(\bar{w};\boldsymbol{\lambda}) - \min_{w} G(w;\bar{\boldsymbol{\lambda}}) \leq \frac{\kappa\mathcal{C}}{T}(\log(\gamma+T-1)+1-\log(\gamma+1)) + \frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}}. \tag{27}$$

Denote the right hand side as ν_T. Pick $T \geq \frac{2\kappa C(\gamma-1)}{\nu(\gamma+1)-2\kappa C}$ and $E \geq \frac{2B^2(\nu(\gamma+1)-2\kappa C)(\log(Z+1)+\rho^2)^2}{\nu^2\kappa C(\gamma-1)}$.

$$\nu_T \leq \frac{\kappa\mathcal{C}}{T}\frac{\gamma+T-1}{\gamma+1} + \frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}} \tag{28}$$

$$\leq \frac{\nu}{2} + \frac{B(\log(Z+1)+\rho^2)}{\sqrt{ET}} \tag{29}$$

$$\leq \frac{\nu}{2} + \frac{B(\log(Z+1)+\rho^2)\sqrt{\nu(\gamma+1)-2\kappa C}}{\sqrt{2\kappa C(\gamma-1)E}} \tag{30}$$

$$\leq \frac{\nu}{2} + \frac{\nu}{2} \tag{31}$$

$$= \nu \tag{32}$$

□

APPENDIX B
PROOF FOR THEOREM IV.9

We first introduce a few lemmas necessary for the proof of Theorem IV.9.

Lemma B.1. *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

$$|\mathcal{F}(h) - F(h)| \leq 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(1/\delta)} \tag{33}$$

Proof for lemma B.1. Lemma B.1 directly follows Theorem 2 in Mohri et al. [17] with $\lambda_k = \frac{1}{K}$. □

Lemma B.2 (Lemma 1 in Agarwal et al. [9])**.** *Let $(\bar{w}, \bar{\boldsymbol{\lambda}})$ is a ν-approximate saddle point, then*

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

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

Lemma B.3 (Lemma 2 in Agarwal et al. [9])**.** *For any w such that $\mathbf{r}(w) \leq \mathbf{0}_Z$, $F(\bar{w}) \leq F(w) + 2\nu$.*

Lemma B.4 (Generation for BGL)**.** *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 **all** $a \in A$, with probability $1-\delta$, we have*

$$|\mathfrak{r}_a(h) - \mathbf{r}_a(h)| \leq 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_a}\sqrt{\frac{K}{2}\log(|A|/\delta)} \tag{35}$$

Lemma B.5 (Generation for CBGL)**.** *Let*

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

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

$$|\mathfrak{r}_a(h) - \mathbf{r}_a(h)| \leq 2\mathfrak{R}_a(\mathcal{H}) + \frac{M}{m_{a,y}}\sqrt{\frac{K}{2}\log(|A||Y|/\delta)} \tag{36}$$

Denote the right hand side of Lemma B.4 and B.5 for constraint j to be $Gen_{r,j}(\delta)$.

Proof for lemma IV.8. Note that

$$\begin{aligned}
F(\bar{w}) + B\max_{z \in Z} \mathbf{r}_z(\bar{w})_+ - \nu &\leq F(\bar{w}) + \bar{\boldsymbol{\lambda}}^T \mathbf{r}(\bar{w}) && (37)\\
&= G(\bar{w}, \bar{\boldsymbol{\lambda}}) && (38)\\
&\leq \min_{w} G(w, \bar{\boldsymbol{\lambda}}) + \nu && (39)\\
&\leq G(w^*, \bar{\boldsymbol{\lambda}}) + \nu && (40)\\
&= F(w^*) + \bar{\boldsymbol{\lambda}}^T \mathbf{r}(w^*) + \nu && (41)\\
&\leq F(w^*) + \nu. && (42)
\end{aligned}$$

Therefore, we have

$$F(\bar{w}) \leq F(w^*) + 2\nu. \tag{43}$$

Hence,

$$\begin{aligned}
B\max_{z \in Z} \mathbf{r}_z(\bar{w})_+ &\leq F(w^*) - F(\bar{w}) + 2\nu && (44)\\
&\leq M + 2\nu. && (45)
\end{aligned}$$

□

Note that Lemma IV.8 tells us when there exists a solution for Problem 3, 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 outputs a model when Problem 3 has a solution.

Now we provide a proof of Theorem IV.9.

Proof for Theorem IV.9. When there exists a solution to Problem 3: w^*, by Lemma B.1, IV.8, we have with probability $1-\delta/2$

$$\mathcal{F}(\bar{w}) \leq F(\bar{w}) + 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(2/\delta)} \tag{46}$$

$$\leq F(w^*) + 2\nu + 2\mathfrak{R}_m(\mathcal{H}) + \frac{M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(2\delta)} \tag{47}$$

$$\leq \mathcal{F}(w^*) + 2\nu + 4\mathfrak{R}_m(\mathcal{H}) + \frac{2M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(2/\delta)}. \tag{48}$$

Combined with Lemma B.4, B.5, and IV.8, we have for all r_j that encodes a fairness constraint, with probability $1-\delta/2$

$$\mathfrak{r}_j(\bar{w}) \leq \mathbf{r}_j(\bar{w}) + Gen_{r,j}(\delta/2) \tag{49}$$

$$\leq \frac{M+2\nu}{B} + +Gen_{r,j}(\delta/2) \tag{50}$$

Therefore, Theorem IV.9 holds with probability $1-\delta$ in this case.

When there doesn't exist a solution to Problem 3, Algorithm 1 outputs $\bar{w}$ only when $\max_{a\in A}\mathbf{r}_a(\bar{w}) \leq \frac{M+2\nu}{B}$. In certain scenarios, we are still able to obtain

$$\mathfrak{r}_a(\bar{w}) \leq \frac{M+2\nu}{B} + Gen_{r,j}(\delta/2) \tag{51}$$

by applying Lemma IV.8. Since w^* doesn't exist, the following holds vacuously:

$$\mathcal{F}(\bar{w}) \leq \mathcal{F}(w^*) + 2\nu + 4\mathfrak{R}_m(\mathcal{H}) + \frac{2M}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{2m_k}\log(2/\delta)} \tag{52}$$

Therefore, Theorem IV.9 holds for both cases.

□

APPENDIX C
ALGORITHM WITH EXAMPLE-LEVEL PRIVACY

In cross-silo federated learning, it is common to ensure privacy through example-level differential privacy, as the private entities are typically examples in each silo (e.g., patients in a hospital, clients in a bank) [39]. We provide an extension of our PFFL method with example-level privacy in Algorithm 2. Our implementation is based off of DP-SGD [40]. Empirically, we show the privacy-fairness trade-off on two datasets in Figure 7. Here we observe that we can provide differential privacy with our fair FL framework, noting that the fairness guarantee is reduced as we require more privacy (smaller ϵ). Following standard private SGD proof, we also provide privacy guarantee to our algorithm 2 in Theorem C.1. While not the focus of this work, further studying relationships between privacy and fairness in the setting of federated learning is an interesting research question enabled by our framework.

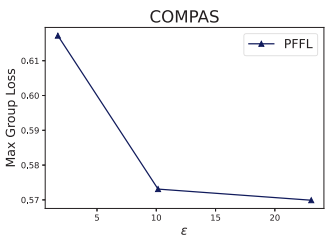

Fig. 7: Given fixed privacy budget ε, we pick the model that yields the lowest max group loss (best fairness). For both datasets, as we require for more privacy (smaller ε), it becomes harder to ensure the utility of the worst performing group.

Theorem C.1. *Assume $M \geq C_r + \zeta$, and f_k is L-Lipschitz such that $L \geq C_w$. Further sssume there are in total n samples and local update subsampling rate is q for all the clients, there exists constants c_1, c_2, c_3, c_4 such that for any $\epsilon < \max\{c_1 ET, c_2 E\}$, Algorithm 2 is $(\epsilon, \delta)-$ example-level differentially private if we choose*

$$\sigma_w = \frac{c_3 L \sqrt{ET \log(1/\delta)}}{n\epsilon} \tag{53}$$

$$\sigma_{\mathbf{r}} = \frac{c_4 \sqrt{2} M \sqrt{E \log(1/\delta)}}{n\epsilon} \tag{54}$$

Lemma C.2. *Let the $g_r(D) = [\sum_k \tilde{\mathbf{r}}_{1,k}, \cdots, \sum_k \tilde{\mathbf{r}}_{Z,k}]$ be the gradient of $\boldsymbol{\lambda}$. Then the ℓ_2 sensitivity of the gradient ascent step $\Delta_2 g_r$ satisfies:*

$$\Delta_2 g_r = \sqrt{2} C_r \tag{55}$$

Proof for Lemma C.2.

$$\begin{aligned}
\Delta_2 g_r &= \sup_{\text{adjacent } D, D'} \|g_r(D) - g_r(D')\|_2 \\
&= \sup_{i,j} \sqrt{\left(\sum_k \tilde{\mathbf{r}}_{i,k} - \sum_k \tilde{\mathbf{r}}'_{i,k}\right)^2 + \left(\sum_k \tilde{\mathbf{r}}_{j,k} - \sum_k \tilde{\mathbf{r}}'_{j,k}\right)^2}
\end{aligned}$$

Assume the supremum is achieved by swapping (x, y, i) to some (x', y', j), then we can rewrite

$$\left(\sum_k \tilde{\mathbf{r}}_{i,k} - \sum_k \tilde{\mathbf{r}}'_{i,k}\right)^2 = \left(\left(\frac{1}{m_i} - \mathbf{1}_{m_i \neq 1} \frac{1}{m_i - 1}\right) S_i + \frac{1}{m_i} \ell(w; x, y, i)\right)^2$$

$$\left(\sum_k \tilde{\mathbf{r}}_{j,k} - \sum_k \tilde{\mathbf{r}}'_{j,k}\right)^2 = \left(\left(\frac{1}{m_j + 1} - \mathbf{1}_{m_j \neq 0} \frac{1}{m_j}\right) S_j + \frac{1}{m_j + 1} \ell(w; x', y', j)\right)^2$$

where $S_i = \sum_{\substack{a_l = i \\ x_l, y_l \neq x, y}} \ell(w; x_l, y_l, i)$, $S_j = \sum_{\substack{a_l = j \\ x_l, y_l \neq x', y'}} \ell(w; x_l, y_l, j)$. Consider the case where $m_i = 1$ and $m_j = 0$, we have $S_i = S_j = 0$. Hence, $\|g_r(D) - g_r(D')\|_2 = \sqrt{2} C_r$ Consider the case where $m_i > 1$ and $m_j > 0$, we have

$$\begin{aligned}
\|g_r(D) - g_r(D')\|_2 &= \sqrt{\left(\left(\frac{1}{m_i} - \frac{1}{m_i - 1}\right) S_i + \frac{C_r}{m_i}\right)^2 + \left(\left(\frac{1}{m_j + 1} - \frac{1}{m_j}\right) S_j + \frac{C_r}{m_j + 1}\right)^2} \\
&\leq \sqrt{\left(\frac{m_i - 2}{m_i (m_i - 1)} C_r\right)^2 + \left(\frac{m_j - 1}{m_j (m_j + 1)} C_r\right)^2} \\
&\leq \sqrt{2\left((3 - 2\sqrt{2}) C_r\right)^2} \\
&= (3\sqrt{2} - 4) C_r
\end{aligned}$$

Note this is less than $\sqrt{2} C_r$. Hence the supremum is achieved when $m_i = 1, m_j = 0$. Therefore $\Delta_2 g_r = \sqrt{2} C_r$. □

Algorithm 2 PFFL-P: Provably Fair Federated Learning with Example-Level Privacy

Input: T, η_w, η_θ,σ_w, $\sigma_{\mathbf{r}}$, C_w, C_r, w_0, M, ν, B, β, ζ, $\theta^0 = \mathbf{0}$, $\bar{w} = \mathbf{0}$
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 client updates its weight w_k for J iterations

$$h_k(w^t, x_{i,k}) = f_k(w^t, x_{i,k}) + \sum_a \mathbf{1}_{a_i,k=a}\boldsymbol{\lambda}_a \cdot \frac{1}{m_a} f_k(w_k^t, x_{i,k})$$

$$w_k^{t+1} = w_k^t - \eta_w \frac{1}{N_k} \cdot \left(\sum_{i=1}^{N_k} \text{Clip}_{C_w}\left(\nabla_{w_k^t} \left(h_k(w^t, x_{i,k}) \right) \right) + \mathcal{N}(0, \sigma_w^2 C_w^2 I) \right)$$

      Compute private $\mathbf{r}_{a,k}^t$:

$$\tilde{\mathbf{r}}_{a,k}^t = \frac{1}{m_a} \sum_{a_{i,k}=a} \text{Clip}_{C_r}(f_k(w_k^t, x_{i,k})) + \mathcal{N}(0, \sigma_{\mathbf{r}}^2 C_r^2 I)$$

      Each client sends $g_k^{t+1} = w_k^{t+1} - w_k^t$ and $\tilde{\mathbf{r}}_{a,k}^t$ 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 θ which would be later used to update the dual variable $\boldsymbol{\lambda}$

$$\theta_a^{(i+1)} = \theta_a^i + \eta_\theta \sum_k \tilde{\mathbf{r}}_{a,k}^t$$

end for
Server updates $\bar{w} \leftarrow \frac{1}{ET}\bar{w}$
Output $\bar{w}$ if $\max_a \mathbf{r}_a \le \frac{M+2\nu}{B}$, and null otherwise.

Now we know the gradient descent step has sensitivity $C_w \le L$ and noise multiplier σ_w. The gradient ascent step has sensitivity $\sqrt{2}C_r \le \sqrt{2}M$ by Lemma C.2 and noise multiplier σ_r. Further note that we perform gradient descent step with ET steps and gradient ascent step with E steps. Following proof from Theorem 1 in Yang et al. [41], Theorem C.1 holds.

APPENDIX D
FULL EXPERIMENTAL RESULTS

In this section, we show the results of group 1 / group 2 utility for three aditional datasets: ACS Income (ACS-I), ACS Public Coverage (ACS-P), and COMPAS, comparing other baseline methods. We see that on ACS-I and ACS-P, our method improves the worst group performance without significantly compromising the other group's performance (unlike AFL on ACS-I). While we do not achieve the smallest accuracy difference on these two tasks compared to FedFB, our method achieves better performance on both groups. On COMPAS, we observe that all group fair methods have similar performance, improving the worst group performance compared to FedAvg.

		FedAvg	AFL	FedFair	FairFed	FedFB	**PFFL (ours)**
ACS-I	Group 1 Acc (↑)	.6693	.637	.6619	.6502	.6598	.6684
	Group 2 Acc (↑)	.6344	.7582	.6301	.6248	.6451	.6529
	Acc Difference (↓)	.0349	.1212	.0318	.0254	.0147	.0155
ACS-P	Group 1 Acc (↑)	.6699	.6744	.6682	.6739	.6525	.6774
	Group 2 Acc (↑)	.6306	.633	.6252	.633	.6244	.6371
	Acc Difference (↓)	.0393	.0414	.043	.0409	.0281	.0403
COMPAS	Group 1 Acc (↑)	.7661	.7431	.7431	.7431	.7431	.7431
	Group 2 Acc (↑)	.7094	.7143	.7123	.7143	.7152	.7143
	Acc Difference (↓)	.0567	.0288	.0308	.0288	.0279	.0288

TABLE V: Comparison between PFFL and prior methods in terms of the test performance metric associated with each group on three additional datasets.

APPENDIX E
SCENARIO WHERE EACH CLIENT IS A DISTINCT GROUP

In this work we aim to develop a general group fair federated learning framework applicable to cross-silo federated learning—a scenario where each client/silo contains multiple data points, each belonging to a possibly different underlying group. In the limiting scenario where *all data points* in each silo belong to the *same group* (i.e., we can treat each client/silo as a distinct group), our method can be viewed as a generalized version of AFL [17] where we allow the constraint term to encode different values of ζ. We compare our method with GIFAIR [20] and q-FFL [18], a variant of AFL that encourages providing fair utility performance across all the clients. For fair comparison, we plot the average test accuracy w.r.t the largest client's loss for both methods. Our method achieves comparable results with prior works in this setting, and both methods yield models that are both fairer and more accurate than vanilla FedAvg. As discussed in Section II, unlike q-FFL/GIFAIR/AFL, the focus of this work is instead to capture group fairness constraints across clients, where a client contains multiple data points and each data point for a particular client belongs to a protected group.

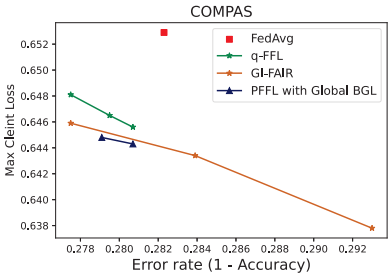

Fig. 8: Comparison between PFFL, GIFAIR, and q-FFL on COMPAS.

APPENDIX F
COMPARISON BETWEEN DIFFERENT SOLVERS

In this section we compare our suggested PFFL solver with the general-purpose saddle point optimization solver (FedAvg-S) proposed in Hou et al. [32]. We refer to FedAvg-S on our objective as PFFL-S. We tune $p \in \{0.25, 0.5, 0.75, 1\}$ (when $p = 1$, it reduces to Peng et al. [42]) and pick the one with the strongest fairness guarantee given same error rate. The results are shown in Figure 9. Our solver is comparable / better than PFFL-S in all settings except when using BGL for the ACS Employment dataset. Our simple solver has the following advantages over FedAvg-S. First, our method supports partial client participation, which is not true in FedAvg-S. Further, FedAvg-S requires an extra hyerparameter p whereas our solver avoids this additional hyperparameter tuning step. While we therefore suggest using the PFFL solver for the BGL objective due to these practical benefits, we note that the fairness/utility results for our BGL objective may be improved even further over existing baselines if a user wishes to optimize over both solvers.

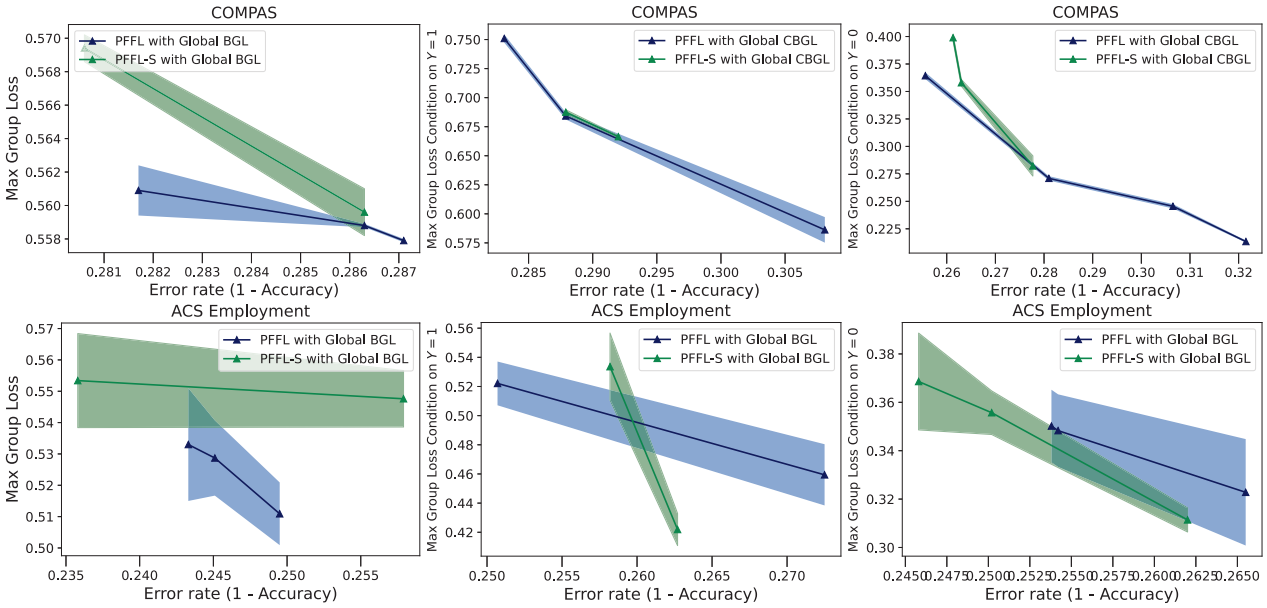

Fig. 9: Comparison between PFFL and PFFL-S on COMPAS and ACS.

APPENDIX G
COMPARISON WITH CENTRALIZED BASELINE

We show the results of comparing our PFFL with BGL method with BGL in centralized setting method in Figure 10 on COMPAS. As expected, centralized BGL is able to produce a model that is both fairer and more accurate than PFFL with Gloabl BGL.

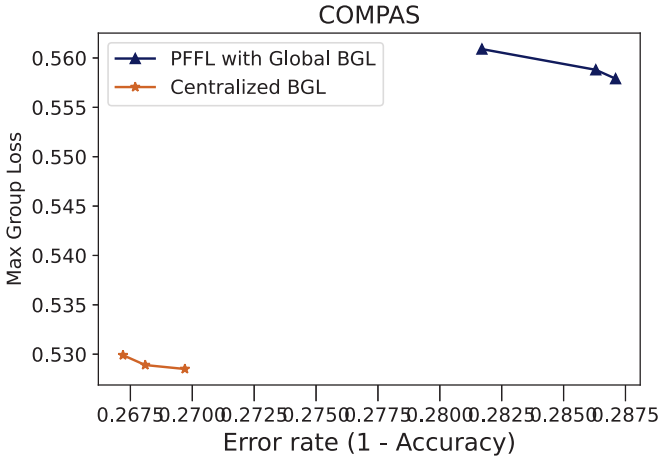

Fig. 10: Comparison between PFFL and using BGL on centralized dataset on COMPAS.