Post-Fair Federated Learning: Achieving Group and Community Fairness in Federated Learning via Post-processing

Anonymous Authors[1]

Abstract

Federated Learning (FL) is a distributed machine learning framework in which a set of local communities collaboratively learn a shared global model while retaining all training data locally within each community. Two notions of fairness have recently emerged as important issues for federated learning: group fairness and community fairness. Group fairness requires that a model's decisions do not favor any particular group based on a set of legally protected attributes such as race or gender. Community fairness requires that global models exhibit similar levels of performance (loss) across all collaborating communities. Both fairness concepts can coexist within an FL framework, but the existing literature has focused on either one concept or the other. This paper proposes and analyzes a post-processing *fair federated learning* (FFL) framework called *post-FFL*. Post-FFL uses a linear program to simultaneously enforce group and community fairness while maximizing the utility of the global model. Because Post-FFL is a post-processing approach, it can be used with existing FL training pipelines whose convergence properties are well understood. Analysis of Post-FFL shows how it can be used to estimate the accuracy lost in simultaneously enforcing group and community fairness. This paper uses post-FFL on real-world datasets to mimic how hospital networks, for example, use federated learning to deliver community health care. The experimental results illustrate that post-FFL simultaneously improves both group and community fairness in Federated Learning. Moreover, it is an effective tool for estimating the accuracy compromised to enhance fairness in Federated Learning.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017) is a distributed machine learning framework that uses data collected from a group of community *clients* to learn a global model that can be used by all clients in the group. The communities served by these clients are formed from sets of remote users (e.g. mobile phones) or organizations (e.g. medical clinics and hospitals) that all share some defining attribute such as a similar geographical location. FL algorithms such as `FedAvg` (McMahan et al., 2017) train the global model in a distributed manner by first having each community client use its local data to train a local model. This local model is then sent to the cloud server who averages these models and sends the averaged model back to the community clients who then retrain that model with their local data. This interaction between the clients and server continues for a several update cycles until it converges on a global model that is agreeable to all clients. While there were no theoretical convergence guarantees with the original FL algorithm (McMahan et al., 2017), subsequent analysis (Smith et al., 2018; Li et al., 2020) did provide theoretical convergence analysis for this FL training pipeline. Since then FL has come to be a dominant framework for distributed machine learning (Kairouz et al., 2021), particularly in smart city (Jiang et al., 2020; Zheng et al., 2022; Qolomany et al., 2020; Pandya et al., 2023) and smart healthcare applications (Rieke et al., 2020; Nguyen et al., 2022; Antunes et al., 2022; Brisimi et al., 2018) .

This paper considers two related notions of fairness relevant to federated learning: Group Fairness and Community Fairness. Group fairness (Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017) is concerned with achieving similar outcomes for groups defined by legally protected (a.k.a. sensitive) attributes such as race or gender. Community fairness (Gross, 2007; 2008) is concerned with the equal allocation of benefits across all communities regardless of their legally protected status. For community fairness, a community may consist of individuals living in the same geographic location. Community fairness, therefore, is more concerned with ensuring that these geographically distinct communities have equal access to resources. Group fairness, on the other hand, requires that all individuals with the

same legally protected attribute receive the same benefits as those outside of the protected group regardless of their membership in these geographically distinct communities or neighborhoods. Both fairness concepts (group vs. community) are relevant to federated learning. This is particularly true in smart healthcare applications where a physician's decisions should not be influenced by factors such as age, race, or gender (Parsa-Parsi, 2017) and yet city leaders want to ensure that all geographically distinct neighborhoods perceive they have the same accessibility to adequate health care. Whether one can balance these two fairness concepts in an FL platform and what might be the cost of attaining such balance is the main topic of this paper.

Several recent papers have proposed methods for achieving either group or community fairness in federated learning. The methods fall into three categories: pre-processing, in-processing, and post-processing. Pre-processing techniques achieve model fairness by modifying the data set used to train the model. This may be done by weighting the training samples as described in (Abay et al., 2020). Pre-processing techniques, however, cannot simultaneously address group and community fairness. In-processing techniques typically modify the federated learning framework's optimzation algorithms. Current approaches either employ dynamic aggregation weights (Yue et al., 2023; Ezzeldin et al., 2023; Chu et al., 2021; Rodríguez-Gálvez et al., 2021; Lyu et al., 2020; Li et al., 2019) or use adversarial training (Du et al., 2021; Mohri et al., 2019). These approaches, however, complicate the existing FL training pipeline and lack formal convergence guarantees. Post-processing, on the other hand, uses models selected by an existing training to generate a randomized model that achieves fairness (Hardt et al., 2016; Fish et al., 2016; Menon & Williamson, 2017; Pleiss et al., 2017; Chzhen et al., 2019; Denis et al., 2021; Zhao & Gordon, 2022; Zeng et al., 2022; Xian et al., 2023). This prior post-processing work, however, does not consider a federated learning (FL), which is the subject of this paper.

This paper's novel contributions are:

- the development of a post-processing FL framework (post-FFL) that simultaneously enforces group fairness and community fairness,
- results that characterize when post-processing can simultaneously achieve group and community fairness for a given group of communities,
- results that allow one to evaluate the model's accuracy lost in achieving group and community fairness,
- and finally the experiment results on a real-world dataset, which show that our framework outperforms existing baselines in both group fairness and community fairness improvement, as well as in communication efficiency and computation cost.

2. Preliminary Definitions

This section provides a statistical interpretation of group and community fairness that allows us to address fairness issues in the federated learning of models that predict outcomes for individuals in a group of communities. The community group is a collection of geographically distinct human communities. Each community is a *client* that uses the local data it has on its inhabitants to select a local model that predicts health outcomes for a given inhabitant. All community clients send their local models to a global *server* who then aggregate the models into a global model. The resulting global model, however, may not be fair either with respect to group or community notions defined below. The main problem is to find a way to *transform* the global model into a *fair global model*. Since this transformation is done after the FL pipeline has selected the global model, this is a post-processing approach to achieving fairness.

Notational Conventions: This paper will denote random variables using upper case letters, X, and lower case letters, x will denote instances of those random variables. A random variable's distribution will be denoted as F_X and an instance, x, drawn from that distribution will be denoted as $x \sim F_X$. Bold face lower case symbols will be reserved for vectors and bold face upper case symbols will be reserved for matrices.

To formalize our statistical setup, we first need to define the notion of a **community group**.

Definition 2.1. A **community group** consists of K geographically distinct communities that we formally represent as a tuple of jointly random variables, $D = (X, A, C, Y)$ with probability distribution $F_D : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \times \mathcal{Y} \to [0, 1]$. An instance of the community group, (x, a, c, y), is called an **individual** where $x \in \mathcal{X}$ is the individual's **private data vector**, $a \in \mathcal{A} = \{0, 1\}$ denotes the individual's **protected sensitive attribute**, and $c \in \mathcal{C} = \{1, 2, \ldots, K\}$ denotes which **community** the individual belongs to. The other value, $y \in \mathcal{Y} = \{0, 1\}$ denotes the individual's *qualified outcome*.

The variables in definition 2.1 have concrete interpretations in a community health application. Each community is a geographically distinct neighborhood served by a single health clinic. For an individual $(x, a, c, y) \sim F_D$, the variable x represents that individual's private health data, a may represent a protected attribute such as race or gender, c is the individual's local health clinic. Finally y represents the whether or not the individual is ill and needs to access medical resources to treat that illness.

We are interested in selecting an **outcome predictor**, $\widehat{Y} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$ for community group D such that for any individual $(x, a, c, y) \sim F_D$ we have $\widehat{Y}(x, a, c) = y$ with a high probability. In particular, let $(1 - \Delta) \in (0, 1)$ denote

a specified accuracy level, then the outcome predictor is Δ-accurate if $\Pr_D\left\{\widehat{Y}(X,A,C)=Y\right\} \geq 1-\Delta$. With these definitions and notational conventions we can now formalize a specific notion of group fairness known as *equal opportunity* (Hardt et al., 2016).

Definition 2.2. . The outcome predictor $\widehat{Y}: \mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ for community group D satisfies **equal opportunity** if and only if

$$\begin{aligned}&\Pr_D\left\{\widehat{Y}(X,A,C)=1\,|\,Y=1,A=1\right\}=\\&\Pr_D\left\{\widehat{Y}(X,A,C)=1\,|\,Y=1,A=0\right\}\end{aligned} \tag{1}$$

Definition 2.2 asserts that the probability of the outcome predictor correctly predicting a positive outcome for an individual (x,a,c,y) from D who qualifies for positive outcome (i.e. $y=1$) is independent of the individual's protected attribute a. The following definition provides a statistical characterization of **community fairness** that is similar to the concept of *fair resource allocation* in (Li et al., 2019).

Definition 2.3. The outcome predictor $\widehat{Y}: \mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ for community group D satisfies **community fairness** if and only if for any $j,k\in\mathcal{C}$, we have

$$\begin{aligned}&\Pr_D\left\{\widehat{Y}(X,A,C)=Y\,|\,C=j\right\}=\\&\Pr_D\left\{\widehat{Y}(X,A,C)=Y\,|\,C=k\right\}\end{aligned} \tag{2}$$

Definition 2.3 asserts that the probability of the outcome predictor correctly predicting an individual's qualified outcome is independent of which community the individual belongs to.

This paper develops a post-processing FL algorithm that selects an outcome predictor $\widetilde{Y}: \mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ that satisfies both *community fairness* and *group fairness*. If a Δ-accurate predictor exists that achieves community and group fairness on community group D, then we say the community group is Δ-equalizable. This paper will also establish necessary conditions for a community group to be Δ-equalizable.

3. Achieving Group and Community Fairness

Let $D=(X,A,C,Y)$ be a community group and consider the loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to\{0,1\}$ that takes values

$$\ell(\widetilde{y},y)=\mathbb{1}(\widetilde{y}\neq y) \tag{3}$$

for any $\widetilde{y},y\in\mathcal{Y}$, where $\mathbb{1}(\cdot)$ is the indicator function. A **fair outcome predictor** is any map $\widetilde{Y}:\mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ that satisfies the equal opportunity equation (1) and community fairness equation (2) with respect to community group D. The following proposition asserts that if an outcome predictor $\widetilde{Y}$ satisfies the following optimization problem in equation (4), then $\widetilde{Y}$ must be a fair outcome predictor with respect to community group D.

Proposition 3.1. *Consider the community group, $D=(X,A,C,Y)$, and the binary loss function, ℓ, in equation (3). If the outcome predictor $\widetilde{Y}:\mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ satisfies the following optimization problem*

$$\begin{aligned}&\textit{minimize} && \mathbb{E}_D\left[\ell(\widetilde{Y}(X,A,C),Y)\right]\\&\textit{with respect to} && \widetilde{Y}:\mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}\\&\textit{subject to} && \Pr_D(\widetilde{Y}(X,A,C)=1\,|\,Y=1,A=0)\\&&&=\Pr_D(\widetilde{Y}(X,A,C)=1\,|\,Y=1,A=1),\\&&& \Pr_D(\widetilde{Y}(X,A,C)=Y\,|\,C=j)\\&&&=\Pr_D(\widetilde{Y}(X,A,C)=Y\,|\,C=k),\forall j,k\in\mathcal{C}\end{aligned} \tag{4}$$

*then $\widetilde{Y}$ is a **fair outcome predictor**.*

Proof: Any $\widetilde{Y}$ that solves optimization problem must satisfy the given constraints. Since these constraints are equation (1) and (2) for equal opportunity and community fairness, respectively, the outcome predictor must also be fair with respect to community group D. $\diamondsuit$

It is not yet clear if the optimization problem in equation (4) actually has a solution. To obtain conditions for the existence of a fair outcome predictor, we will first show that equation (4) can be recast as a linear program. The existence of a fair outcome predictor is then equivalent to that linear program having non-negative solutions.

Let $\widehat{Y}:\mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ be an *optimal* outcome predictor that minimizes the expected value of the indicator loss function in equation (3) with respect to community group D. For notational convenience we will drop the arguments on the outcome predictors so we write $\widehat{Y}(X,A,C)$ (or $\widetilde{Y}(X,A,C)$) as $\widehat{Y}$ (or $\widetilde{Y}$). For convenience we introduce the following notational conventions for the optimal outcome predictor's joint probability of false or true positives and negatives:

$$\begin{aligned}\mathrm{FN}^{ac}&=\Pr_D\left\{\widehat{Y}=0,Y=1,A=a,C=c\right\}\\\mathrm{TN}^{ac}&=\Pr_D\left\{\widehat{Y}=0,Y=0,A=a,C=c\right\}\\\mathrm{FP}^{ac}&=\Pr_D\left\{\widehat{Y}=1,Y=0,A=a,C=c\right\}\\\mathrm{TP}^{ac}&=\Pr_D\left\{\widehat{Y}=1,Y=1,A=a,C=c\right\}\end{aligned} \tag{5}$$

We will also find it convenient to define the following statistics that can be from the group community, D.

$$\begin{aligned}p_c&=\Pr_D(C=c)\\\alpha&=\Pr_D(Y=1\,,\,A=0)\\\beta&=\Pr_D(Y=1\,,\,A=1)\end{aligned} \tag{6}$$

p_c is the probability of random individual being in community c, α is the probability of random individual being qualified and non-sensitive, β is the probability of random individual being qualified and sensitive.

Finally, the variables we will use to characterize our fair outcome predictor, $\widetilde{Y}$, will be

$$z_j^{ac} \quad = \quad \Pr_D\left\{\widetilde{Y} = \widehat{Y} \,|\, \widehat{Y} = j, A = a, C = c\right\} \quad (7)$$

So z_j^{ac} is the probability that the fair predictor's outcome, $\widetilde{Y}$, equals that of the optimal predictor's outcome, $\widehat{Y}$, for an individual, (x, a, c, y), for which the optimal predictor's output is $j \in \mathcal{Y}$, the sensitive attribute is $a \in \mathcal{A}$ and the community label is $c \in \mathcal{C}$.

Proposition 3.2. *(Appendix B.1) Let $\mathbf{z} \in \mathbb{R}^{4K}$ satisfy the following linear program*

$$\begin{array}{ll} \textit{minimize:} & \mathbf{c}^T\mathbf{z} \\ \textit{with respect to:} & \mathbf{z} \in \mathbb{R}^{4K} \\ \textit{subject to:} & \mathbf{A}\mathbf{z} = \mathbf{b} \\ & 0 \le \mathbf{z} \le 1 \end{array} \quad (8)$$

where the parameters of the linear program, $\mathbf{A} \in \mathbb{R}^{(K+1)\times 4K}$, $\mathbf{c} \in \mathbb{R}^{4K}$, and $\mathbf{b} \in \mathbb{R}^{K+1}$, are constructed using the statistics defined in (5) and (6). The concrete representation of the linear program is in Appendix A.

Let the solution of the linear program $\mathbf{z}$ be:

$$\mathbf{z}^T \quad = \quad \left[\begin{array}{cccc} \mathbf{z}_1^T & \mathbf{z}_2^T & \cdots & \mathbf{z}_K^T \end{array}\right]$$

with

$$\mathbf{z}_i^T \quad = \quad \left[\begin{array}{cccc} z_0^{0i} & z_1^{0i} & z_0^{1i} & z_1^{1i} \end{array}\right]$$

Then the outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$ taking values

If: $\widehat{Y}(x, a, c) = 0, A = a, C = c$:

$$\widetilde{Y}_{\widehat{Y},\mathbf{z}}(x, a, c) = \begin{cases} 0 & \textit{with probability } z_0^{ac} \\ 1 & \textit{with probability } 1 - z_0^{ac} \end{cases}$$

If: $\widehat{Y}(x, a, c) = 1, A = a, C = c$:

$$\widetilde{Y}_{\widehat{Y},\mathbf{z}}(x, a, c) = \begin{cases} 1 & \textit{with probability } z_1^{ac} \\ 0 & \textit{with probability } 1 - z_1^{ac} \end{cases}$$

is a fair outcome predictor.

Remark 3.3. It is worth noting that the optimization shown in (8) can incorporate a more relaxed version of equal opportunity and community fairness constraints:

$$|\Pr_D(\tilde{Y} = 1|Y = 1, A = 0) - \Pr_D(\tilde{Y} = 1|Y = 1, A = 1)| \le \epsilon$$
$$\forall k \in \mathcal{C}, |\Pr(\widetilde{Y} \ne Y|C = k) - \frac{1}{K}\sum_{c=1}^{K}\Pr(\widetilde{Y} \ne Y|C = c)| \le \delta \quad (9)$$

In such cases, the linear program with the relaxed fairness constraints is:

$$\begin{array}{ll} \text{minimize:} & \mathbf{c}^T\mathbf{z} \\ \text{with respect to:} & \mathbf{z} \in \mathbb{R}^{4K} \\ \text{subject to:} & \mathbf{b} - \boldsymbol{\epsilon} \le \mathbf{A}\mathbf{z} \le \mathbf{b} + \boldsymbol{\epsilon} \\ & 0 \le \mathbf{z} \le 1 \end{array} \quad (10)$$

where,

$$\boldsymbol{\epsilon}^T \quad = \quad \left[\begin{array}{cccc} \epsilon & \delta & \cdots & \delta \end{array}\right]$$

By resolving the linear program, with different $\boldsymbol{\epsilon} \in \mathbb{R}^{(K+1)}$, we can precisely control the degree of community fairness and group fairness, setting $\boldsymbol{\epsilon} = \mathbf{0}$ gives us the outcome predictor that strictly satisfies equal opportunity and community fairness.$\diamondsuit$

We can write the linear program (8) in the *standard form* by introducing a set of slack variables $\mathbf{s} \in \mathbb{R}^{4K}$:

$$\mathbf{s}^T \quad = \quad \left[\begin{array}{cccc} \mathbf{s}_1^T & \mathbf{s}_2^T & \cdots & \mathbf{s}_K^T \end{array}\right]$$

with

$$\mathbf{s}_i^T \quad = \quad \left[\begin{array}{cccc} s_0^{0i} & s_1^{0i} & s_0^{1i} & s_1^{1i} \end{array}\right]$$

The variables we need to solve for in linear program (8) represents the probabilities, thus, $0 \le z_j^{ac} \le 1$. It is equivalent to:

$$\begin{array}{c} z_j^{ac} + s_j^{ac} = 1 \\ z_j^{ac}, s_j^{ac} \ge 0 \end{array} \quad (11)$$

Combine the linear program (8) with (11), the standard form of the linear program (8) is:

$$\begin{array}{ll} \text{minimize:} & \bar{\mathbf{c}}^T\bar{\mathbf{z}} \\ \text{with respect to:} & \bar{\mathbf{z}} \in \mathbb{R}^{8K} \\ \text{subject to:} & \bar{\mathbf{A}}\bar{\mathbf{z}} = \bar{\mathbf{b}} \\ & \bar{\mathbf{z}} \ge 0 \end{array} \quad (12)$$

with

$$\begin{aligned} \bar{\mathbf{c}}^T &= \left[\begin{array}{cc} \mathbf{c}^T & \mathbf{0} \end{array}\right] \in \mathbb{R}^{8K} \\ \bar{\mathbf{z}}^T &= \left[\begin{array}{cc} \mathbf{z}^T & \mathbf{s}^T \end{array}\right] \in \mathbb{R}^{8K} \\ \bar{\mathbf{A}} &= \left[\begin{array}{cc} \mathbf{A} & \mathbf{0} \\ \mathbf{I} & \mathbf{I} \end{array}\right] \in \mathbb{R}^{(5K+1)\times 8K} \\ \bar{\mathbf{b}}^T &= \left[\begin{array}{cc} \mathbf{b}^T & \mathbf{1}_{4K}^T \end{array}\right] \in \mathbb{R}^{5K+1} \end{aligned}$$

where $\mathbf{1}_{4K} \in \mathbb{R}^{4K}$ is all 1 vector.

Theorem 3.4. *(Appendix B.2) The linear program (12) always has non-negative solutions.*

Theorem 3.4 indicates that the linear program (8), which is equivalent to (12), always has a solution. Thus, there always exists fair outcome predictors satisfying both equal opportunity and community fairness. We next demonstrate the necessary conditions for the existence of a Δ-accurate fair outcome predictor. An outcome predictor: $\widetilde{Y} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$ is a Δ-accurate fair outcome predictor if:

$$\begin{aligned} &\Pr_D(\widetilde{Y} \neq Y) \leq \Delta \\ &\Pr_D(\widetilde{Y} = 1 | Y = 1, A = 0) = \Pr_D(\widetilde{Y} = 1 | Y = 1, A = 1) \\ &\forall k \in \mathcal{C}, \Pr_D(\widetilde{Y} \neq Y | C = k) = \frac{1}{K} \sum_{c=1}^{K} \Pr_D(\widetilde{Y} \neq Y | C = c) \end{aligned} \tag{13}$$

Theorem 3.5. *(Appendix B.3) The three conditions in (13) are impossible to hold simultaneously if* $\Delta < -\|\bar{\mathbf{c}}\|_\infty \frac{\|\bar{\mathbf{A}}^T \bar{\mathbf{b}}\|_2}{\underline{\sigma}^2} + \sum_{c=1}^{K} (\mathrm{TN}^{0c} + \mathrm{TP}^{0c} + \mathrm{TN}^{1c} + \mathrm{TP}^{1c})$, *where,* $\underline{\sigma}$ *is the smallest singular value of the matrix* $\bar{\mathbf{A}}$.

Theorem 3.5 establishes a necessary condition for the existence of an Δ-accurate fair outcome predictor, based on the statistics of the data represented by $\bar{\mathbf{A}}, \bar{\mathbf{b}}, \bar{\mathbf{c}}$.

Theorem 3.6. *(Appendix B.4) Let* $\mathbf{z} \in \mathbb{R}^{4K}$ *be the solution of the linear program (10) with a predefined* $\boldsymbol{\epsilon}$*, then the minimum accuracy we lose for improving both community fairness and group fairness under post-FFL is* $\mathbf{c}^T(\mathbf{z} - \mathbf{1}_{4k})$, *where,* $\mathbf{1}_{4K}$ *is all* 1 *vector.*

Theorem 3.6 above can serve as a tool for evaluating the accuracy we lose for improving group fairness and community fairness

4. Post-FFL: Fair Outcome Predictor in FL

This section shows how the linear program (8) can be used within a federated learning framework to construct a fair outcome predictor. An overview of post-FFL is shown in Fig.1. We provide the concrete training steps of post-FFL:

1. **Training an Optimal Outcome Predictor using *FedAvg***: The server and communities collaboratively train an optimal predictor $\widehat{Y} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$ using the *FedAvg* algorithm (McMahan et al., 2017). The participating community trains a local model at time t and sends the local model's parameters, $\boldsymbol{\theta}_c^t \in \mathbb{R}^n$, to the global server. The global server then aggregates the local models into a global model parameter, $\boldsymbol{\theta}^t = \sum_{c=1}^{K} p_c \boldsymbol{\theta}_c^t$, and sends it back to the local communities. This iterative process continues until the parameters, $\boldsymbol{\theta}$, of the optimal predictor $\widehat{Y}$ are determined. The aggregation weight, $p_c = \Pr_D(C = c)$, is estimated from the dataset as shown in (14).

$$p_c = \frac{\text{number of samples in community } c}{\text{number of total samples}} \tag{14}$$

2. **Local Prediction and Probability Calculation**: Each local community generates predictions $\widehat{Y}(X, A, C) \in \mathcal{Y}$, computes local statistics $\Pr_D(\widehat{Y} = y, Y = y', A = a \mid C = c)$ for all $(y, y', a) \in \{0, 1\}^3$ as specified in (15), and then transmits these probabilities to the global server.

$$\Pr_D(\hat{Y} = y, Y = y', A = a \mid C = c) = \frac{\text{number of samples with } (\hat{Y} = y, Y = y', A = a) \text{ in community } c}{\text{number of samples in community c}} \tag{15}$$

3. **Constructing and Solving the Linear Program**: The global server computes the parameters defined in Equations (5) and (6) using the probabilities sent by the communities. The parameters $\mathrm{FN}^{\mathrm{ac}}, \mathrm{TN}^{\mathrm{ac}}, \mathrm{FP}^{\mathrm{ac}}, \mathrm{TP}^{\mathrm{ac}}$ in (5) are computed as:

$$\begin{aligned} &\Pr_D\left\{\widehat{Y} = y, Y = y', A = a, C = c\right\} \\ =&p_c \Pr_D\left\{\widehat{Y} = y, Y = y', A = a \mid C = c\right\} \end{aligned}$$

The parameters α and β in (6) are computed as:

$$\begin{aligned} \alpha &= \Pr_D(Y = 1, A = 0) = \sum_{c=1}^{K} (\mathrm{FN}^{0c} + \mathrm{TP}^{0c}) \\ \beta &= \Pr_D(Y = 1, A = 1) = \sum_{c=1}^{K} (\mathrm{FN}^{1c} + \mathrm{TP}^{1c}) \end{aligned}$$

Using the above parameters, the global server constructs the linear program (8), finds the minimizer $\mathbf{z}$:

$$\mathbf{z}^T = \begin{bmatrix} \mathbf{z}_1^T & \mathbf{z}_2^T & \cdots & \mathbf{z}_K^T \end{bmatrix}$$

and then sends the corresponding minimizer $\mathbf{z}_k^T$ to community k, where $k = 1, 2, \cdots, K$.

4. **Fair Outcome Predictor**: The local community k, $(k = 1, 2, \cdots K)$ employs Algorithm 1 to make fair predictions. The received minimizer $\mathbf{z}_k^T$ indicates the probability that the fair predictor's outcome equals to the optimal predictor's outcome for community k.

Algorithm 1 provides a community dependent randomized function that is used to decide whether to accept or deny the prediction from the optimal model. The output of the optimal model, combined with the randomized function, will yield a fair outcome predictor.

5. Experiments

We conduct experiments on the real-world dataset to demonstrate that our framework is an effective tool for controlling the degree of both fairness and estimating the accuracy lost for improving fairness. It outperforms the existing in communication effencicy and computation cost.

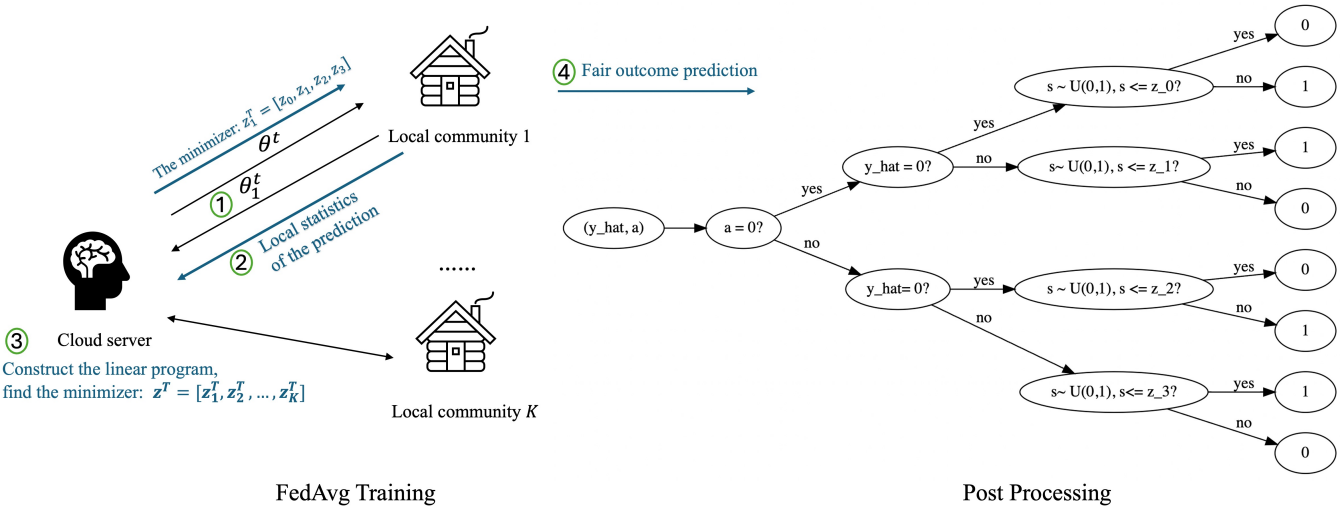

Figure 1. Overview of the proposed post-FFL framework. The black arrow signifies the exchange of local models θ_i^t and global models θ^t and during *FedAvg* model training. The blue arrow depicts the post-processing workflow following *FedAvg*, where local communities forward their local statistics to the global server. The global server constructs a linear program and send the solution back to local communities. Each local community uses a decision tree, as shown on the right, to make fair outcome predictions.

Algorithm 1 Fair Outcome Predictor

Input: The optimal outcome predictor: $\widehat{Y} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$, the community k's corresponding minimizer:
$\mathbf{z}_k^T = \begin{bmatrix} z_0^{0k} & z_1^{0k} & z_0^{1k} & z_1^{1k} \end{bmatrix}$

Output: Fair outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}_k} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$

1. randomly sample $s \sim U(0,1)$, the uniform distribution between (0,1)
2. Construct $\widetilde{Y}_{\widehat{Y},\mathbf{z}_k}(x,a,k)$ as

$$\widetilde{Y}_{\widehat{Y},\mathbf{z}_k}(x,a,k) = \begin{cases} a=0: \begin{cases} 0 & \text{If } (\widehat{Y}=0 \text{ and } s \le z_0^{0k}) \text{ or } (\widehat{Y}=1 \text{ and } s > z_1^{0k}) \\ 1 & \text{If } (\widehat{Y}=0 \text{ and } s > z_0^{0k}) \text{ or } (\widehat{Y}=1 \text{ and } s \le z_1^{0k}) \end{cases} \\ a=1: \begin{cases} 0 & \text{If } (\widehat{Y}=0 \text{ and } s \le z_0^{1k}) \text{ or } (\widehat{Y}=1 \text{ and } s > z_1^{1k}) \\ 1 & \text{If } (\widehat{Y}=0 \text{ and } s > z_0^{1k}) \text{ or } (\widehat{Y}=1 \text{ and } s \le z_1^{1k}) \end{cases} \end{cases}$$

return $\widetilde{Y}_{\widehat{Y},\mathbf{z}_k}$

5.1. Experimental Setup

Dataset. We demonstrate the effectiveness of our framework on two real-world datasets. The datasets we use are the Adult dataset and the Diabetes dataset.

The **Adult** dataset (Asuncion & Newman, 2007) consists of 6 numerical features (age, final weight, education number, etc.) and 8 categorical features (work class, education, gender, race, etc.) and is used to predict whether an individual earns more than 50K/year. We set gender as the sensitive attribute. Following the federated setting in (Li et al., 2020; Mohri et al., 2019), we split the dataset into two communities: one is the PhD community, in which all individuals are PhDs, and the other is the non-PhD community.

The **Diabetes** dataset (Strack et al., 2014) contains 10 numerical features (time in hospital, number of procedures, etc.), 40 binary features (race, gender, age range, admission source, diabetMed, etc.), and is used to predict whether a patient will be readmitted within 30 days. We set the group 'older (aged over 60) African-American females' as the sensitive group. We split the data into 7 communities based on their admission source. The communities 1, 2, 3, 4, 5, 6, and 7 represent samples admitted from the Emergency room, Physician referral, NULL, Transfer from a hospital, Transfer from another healthcare facility, Clinic referral, Transfer from a Skilled Nursing Facility, and Others, respectively.

Evaluation Metrics. We evaluate the model's performance from the following perspectives: (1) Model utility: We use the model's Average Accuracy (Avg-Acc) as a measure of its utility. The Average Accuracy is the weighted average accuracy across all communities, where the weight of each community is determined by its data size. (2) Group

fairness: Group fairness, as defined in 2.2, requires that the true positive rates are the same for sensitive and non-sensitive groups, so we assess group fairness using the Equal Opportunity Difference (EOD). The EOD is defined as the disparity of true positive rates between sensitive and non-sensitive groups. (3) Community fairness: Community fairness, as defined in 2.3, requires that the model has similar accuracy across all communities. We measure community fairness using the Accuracy Disparity (AD), which is defined as the difference in accuracy between the community with the highest accuracy and the one with the lowest.

The full experiments details (including model and hyperparameters) can be found in Appendix C.1

5.2. Fairness of post-FFL

In this section, we verify that the proposed framework can simultaneously enforce group fairness and community fairness within a FL platform. We demonstrate that our framework can result in a relaxed fair outcome predictor. Specifically, we can control the degree of group fairness and community fairness by adjusting the ϵ and δ in the linear program (10) respectively. Smaller ϵ and δ values will lead to a fairer outcome predictor. Setting $(\epsilon = 0, \delta = 0)$ results in a predictor that strictly achieves both group fairness and community fairness. We demonstrate that Theorem (3.6) within our framework allows one to evaluate the accuracy loss incurred while improving group fairness and community fairness.

We report the Avg-Acc, EOD, and AD of the initial Fedavg and after our post-processing with $(\epsilon = 0, \delta = 0)$ in Table 1. We observe that for the UCI Adult dataset, the EOD and AD of post-FFL are 0.016 and 0.012, respectively, and for the Diabetes dataset, they are 0.008 and 0.021. Compared to the initial Fedavg, post-FFL effectively reduces the Equal Opportunity Difference and Accuracy Disparity, demonstrating that post-FFL can enforce group fairness and community fairness simultaneously.

We present the results of Avg-Acc, EOD, and AD for different settings of (ϵ, δ) in the left side of Table 2. Our framework is flexible in that it allows one to choose (ϵ, δ) to tradoff between fairness and global accuracy. With a fixed ϵ, decreasing δ reduces the AD, indicating a model that is fairer with respect to community fairness. Similarly, with a fixed δ, the degree of equal opportunity can be effectively controlled.

We next show that the the average accuracy loss for improving fairness aligned with our theoretical analysis. In the right of Table 2, the empirical accuracy loss is the disparity between the model's accuracy under the initial FedAvg and its accuracy after applying post-FFL adjustments from our experiments. The estimated accuracy loss is calculated using the theoretical result (3.6), which states the accuracy we loss is $\mathbf{c}^T(\mathbf{z} - \mathbf{1_{4K}})$. We observe that the actual accuracy loss for a given degree of fairness closely matches our theoretical estimations. The estimated error is always less than 0.01 across different settings. This proves that the theoretical result (3.6) within our framework is an effective tool for evaluating the accuracy loss associated with improving fairness.

Table 1. The EOD, AD and Avg-Acc of *FedAvg* (McMahan et al., 2017) and our post-FFL

DATASET	FRAMEWORKS	EOD	AD	AVG-ACC
ADULT	FEDAVG	0.106	0.124	0.854
	POST-FFL	0.016	0.012	0.780
DIABETES	FEDAVG	0.057	0.083	0.833
	POST-FFL	0.008	0.021	0.812

5.3. Comparison with other objectives

We compare our post-FFL framework with other objectives. We did not find prior work that tries to simultaneously achieve group and community fairness in a federated setting. The most relevant prior work employs in-processing (rather than post-processing) techniques to achieve either equal opportunity or fair resource allocation (community fairness) in a federated setting. We use *q-FedAvg* (Li et al., 2019) and *FairFed* (Ezzeldin et al., 2023) as baselines for community fairness and group fairness, respectively. We modified the original in-processing techniques to regularize with respect to both fairness concepts. Specifically, we developed *q-FedAvg+FairFed*. The global model parameter is set as $\boldsymbol{\theta}^t = \lambda \boldsymbol{\theta_1}^t + (1-\lambda)\boldsymbol{\theta_2}^t$, with $\boldsymbol{\theta}_1^t$ representing the global model updated by *FairFed*, and $\boldsymbol{\theta}_2^t$ representing the global model updated by *q-FedAvg*. A full description of all baselines is provided in Appendix C.2

We compare our post-FFL approach with the objectives above on the *Adult* dataset using the three evaluation metrics. The results are reported in Table 3. In our experimental setting, *q-FedAvg* improves community fairness but exacerbates the equal opportunity difference. Conversely, *FairFed* improves equal opportunity while worsening community fairness. For *q-FedAvg+FairFed*, increasing λ improves group fairness; however, this comes at the expense of community fairness. All baselines cannot simultaneously improve both group fairness and community in federated learning. Our post-FFL demonstrates the ability to reduce both Equal Opportunity Difference (EOD) and Accuracy Difference (AD) at a minimal level. Therefore, our method outperforms the baselines in improving both group fairness and community fairness.

We show the convergence behavior of *FedAvg*, *FairFed* and

Table 2. Post-FFL with varying (ϵ, δ): The left side of the table illustrates we can adjust ϵ and δ to control the level of group fairness and community fairness. The right side of the table shows that the estimated accuracy loss, as evaluated by Theorem 3.6, closely matches the empirical accuracy loss in experiments.

Dataset	(ϵ, δ)	EOD	AD	Avg-Acc	empirical accuracy loss	estimated accuracy loss	estimated error
Adult	(0.00, 0.00)	0.016	0.012	0.780	0.074	0.074	0.000
	(0.00, 0.02)	0.008	0.053	0.804	0.050	0.056	0.006
	(0.00, 0.04)	0.001	0.091	0.841	0.013	0.012	0.001
	(0.02, 0.00)	0.033	0.014	0.766	0.088	0.095	0.007
	(0.02, 0.02)	0.030	0.051	0.802	0.051	0.056	0.004
	(0.02, 0.04)	0.017	0.090	0.837	0.017	0.018	0.001
	(0.04, 0.00)	0.049	0.016	0.758	0.095	0.096	0.001
	(0.04, 0.02)	0.029	0.056	0.797	0.057	0.057	0.000
	(0.04, 0.04)	0.044	0.094	0.839	0.014	0.018	0.004
Diabetes	(0.00, 0.00)	0.008	0.021	0.812	0.022	0.022	0.000
	(0.00, 0.02)	0.034	0.067	0.830	0.002	0.001	0.001
	(0.00, 0.04)	0.047	0.080	0.831	0.003	0.003	0.000
	(0.02, 0.00)	0.040	0.013	0.815	0.018	0.022	0.004
	(0.02, 0.02)	0.025	0.048	0.830	0.004	0.001	0.003
	(0.02, 0.04)	0.027	0.077	0.832	0.002	0.000	0.002
	(0.04, 0.00)	0.057	0.025	0.815	0.018	0.022	0.004
	(0.04, 0.02)	0.006	0.070	0.831	0.002	0.001	0.001
	(0.04, 0.04)	0.032	0.076	0.831	0.002	0.00	0.002

Table 3. Adult dataset: EOD, AD and Avg-Acc of all objectives.

Objectives	EOD	AD	Avg-Acc
Initial *FedAvg*	0.106	0.124	**0.854**
Our *post-FFL* $(\epsilon = 0, \delta = 0)$	**0.016**	**0.012**	0.780
q-FedAvg	0.337	0.002	0.811
FairFed	0.017	0.362	0.846
q-FedAvg+ FairFed$(\lambda = 0.3)$	0.113	0.040	0.848
q-FedAvg+ FairFed$(\lambda = 0.5)$	0.106	0.128	0.845
q-FedAvg+ FairFed$(\lambda = 0.7)$	0.056	0.168	0.837

q-FedAvg. At each communication round, all methods perform the same amount of local model updates, with each completing one epoch of local updates per community, using identical batch sizes and optimization settings. As show in Table 4, the existing in-processing techniques such as *FairFed* and *q-FedAvg* will lead to slower convergence in terms of communication rounds. In contrast, the post-FFL does not change the convergence behavior of the original FedAvg algorithm. It outperforms in-processing fair federated learning methods in the number of round for model convergence and communication efficiency. The training curves are in Appendix C.3

We also report the time taken to complete one communication round for *FedAvg*, *FairFed*, and *q-FedAvg*. In each communication round, all methods first train all local models for 1 epoch using identical batch sizes and optimization settings, compute the model aggregation weights, and finally aggregate all weights in the global model. We report the average time taken by all objectives to complete one round of global model updates in Table 4. We trained all objectives on our local Linux server with a 16-Core 4.00 GHz AMD RYZEN Threadripper Pro 5955WX Processor. The

Table 4. Adult dataset: Number of communication rounds for convergence and average time taken for completing one round of update of *FedAvg*, *q-FedAvg* and*FairFed*.

Objectives	*FedAvg*	*q-Fedavg*	*FairFed*
# of convergence rounds	$\approx$ **5**	$>$ 1000	$\approx$20
Avg-time for 1 round	**0.631s**	0.639s	16.428s

average time is calculated over 30 communication rounds for each objective. We found that *FairFed* requires much more time to update one round, mainly because its aggregation weights are the mismatch between the global EOD and the local EOD. Calculating the EOD for all communities and the global EOD in every round introduces additional computation. The aggregation weights for *q-FedAvg* are a function of local loss, and those for *FedAvg* are static, so they do not require extra computation. The post-FFL does not change the time required for each round of initial *FedAvg*. It outperforms in-process fair federated learning methods in terms of computation cost, as evidenced by the smallest time required for one round of updates.

6. Conclusion

In this work, we propose the post-FFL framework, which simultaneously achieves group and community fairness in FL. Experiments on real-world datasets demonstrate that post-FFL allows users to adjust the degree of fairness based on their requirements. Post-FFL outperforms existing baselines in fair federated learning in terms of fairness improvement, communication efficiency, and computational cost. It is an effective tool for estimating the accuracy of predicted outcomes when smart city and hospital networks seek to simultaneously achieve group and community fairness.

7. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

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

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., and Eskofier, B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13 (4):1–23, 2022.

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., and Shi, W. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

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

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

Denis, C., Elie, R., Hebiri, M., and Hu, F. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

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

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

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

Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pp. 144–152. SIAM, 2016.

Goldman, A. J. and Tucker, A. W. *2. Polyhedral Convex Cones*, pp. 19–40. Princeton University Press, Princeton, 1957. ISBN 9781400881987. doi: doi:10.1515/9781400881987-003. URL https://doi.org/10.1515/9781400881987-003.

Gross, C. Community perspectives of wind energy in australia: The application of a justice and community fairness framework to increase social acceptance. *Energy policy*, 35(5):2727–2736, 2007.

Gross, C. A measure of fairness: An investigative framework to explore perceptions of fairness and justice in a real-life social conflict. *Human Ecology Review*, 15 (2):130–140, 2008. ISSN 10744827, 22040919. URL http://www.jstor.org/stable/24707597.

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

Jiang, J. C., Kantarci, B., Oktug, S., and Soyata, T. Federated learning in smart city sensing: Challenges and opportunities. *Sensors*, 20(21):6230, 2020.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Lyu, L., Xu, X., Wang, Q., and Yu, H. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive*, pp. 189–204, 2020.

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

Menon, A. K. and Williamson, R. C. The cost of fairness in classification. *arXiv preprint arXiv:1705.09055*, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., and Hwang, W.-J. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

Pandya, S., Srivastava, G., Jhaveri, R., Babu, M. R., Bhattacharya, S., Maddikunta, P. K. R., Mastorakis, S., Piran, M. J., and Gadekallu, T. R. Federated learning for smart cities: A comprehensive survey. *Sustainable Energy Technologies and Assessments*, 55:102987, 2023.

Parsa-Parsi, R. W. The Revised Declaration of Geneva: A Modern-Day Physician's Pledge. *JAMA*, 318(20):1971–1972, 11 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.16230. URL https://doi.org/10.1001/jama.2017.16230.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Qolomany, B., Ahmad, K., Al-Fuqaha, A., and Qadir, J. Particle swarm optimized federated learning for industrial iot and smart city services. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6. IEEE, 2020.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

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

Smith, V., Forte, S., Ma, C., Takáč, M., Jordan, M. I., and Jaggi, M. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(230):1–49, 2018.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., Clore, J. N., et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

Xian, R., Yin, L., and Zhao, H. Fair and optimal classification via post-processing. In *International Conference on Machine Learning*, pp. 37977–38012. PMLR, 2023.

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

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Zeng, X., Dobriban, E., and Cheng, G. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

Zheng, Z., Zhou, Y., Sun, Y., Wang, Z., Liu, B., and Li, K. Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges. *Connection Science*, 34(1):1–28, 2022.

A. The Parameters of LP in Proposition 3.2

The linear program (8) is:

$$\begin{array}{ll} \text{minimize:} & \mathbf{c}^T\mathbf{z} \\ \text{with respect to:} & \mathbf{z} \in \mathbb{R}^{4K} \\ \text{subject to:} & \mathbf{A}\mathbf{z} = \mathbf{b} \\ & 0 \leq \mathbf{z} \leq 1 \end{array}$$

with

$$\begin{aligned} \mathbf{c}^T &= \left[\begin{array}{cccc} \mathbf{c}_1^T & \mathbf{c}_2^T & \cdots & \mathbf{c}_K^T \end{array} \right] \\ \mathbf{z}^T &= \left[\begin{array}{cccc} \mathbf{z}_1^T & \mathbf{z}_2^T & \cdots & \mathbf{z}_K^T \end{array} \right] \end{aligned}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{m}_1^T & \mathbf{m}_2^T & \mathbf{m}_3^T & \cdots & \mathbf{m}_{K-1}^T & \mathbf{m}_K^T \\ -\frac{K-1}{K}\mathbf{n}_1^T & \frac{1}{K}\mathbf{n}_2^T & \frac{1}{K}\mathbf{n}_3^T & \cdots & \frac{1}{K}\mathbf{n}_{K-1}^T & \frac{1}{K}\mathbf{n}_K^T \\ \frac{1}{K}\mathbf{n}_1^T & -\frac{K-1}{K}\mathbf{n}_2^T & \mathbf{n}_3^T & \cdots & \frac{1}{K}\mathbf{n}_{K-1}^T & \frac{1}{K}\mathbf{n}_K^T \\ \frac{1}{K}\mathbf{n}_1^T & \frac{1}{K}\mathbf{n}_2^T & -\frac{K-1}{K}\mathbf{n}_3^T & \cdots & \frac{1}{K}\mathbf{n}_{K-1}^T & \frac{1}{K}\mathbf{n}_K^T \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{K}\mathbf{n}_1^T & \frac{1}{K}\mathbf{n}_2^T & \frac{1}{K}\mathbf{n}_3^T & \cdots & -\frac{K-1}{K}\mathbf{n}_{K-1}^T & \frac{1}{K}\mathbf{n}_K^T \\ \frac{1}{K}\mathbf{n}_1^T & \frac{1}{K}\mathbf{n}_2^T & \frac{1}{K}\mathbf{n}_3^T & \cdots & \frac{1}{K}\mathbf{n}_{K-1}^T & -\frac{K-1}{K}\mathbf{n}_K^T \end{bmatrix}$$

$$\mathbf{b}^T = \left[\begin{array}{cccc} \displaystyle\sum_{c=1}^{K}\left(\frac{\mathrm{FN}^{1c}}{\beta} - \frac{\mathrm{FN}^{0c}}{\alpha}\right) & \displaystyle\frac{1}{K}\sum_{c=1}^{K}(b_1 - b_c) & \displaystyle\frac{1}{K}\sum_{c=1}^{K}(b_2 - b_c) & \cdots \quad \displaystyle\frac{1}{K}\sum_{c=1}^{K}(b_K - b_c) \end{array} \right]$$

with

$$\begin{aligned} \mathbf{c}_i^T &= \left[\begin{array}{cccc} (\mathrm{FN}^{0i} - \mathrm{TN}^{0i}) & (\mathrm{FP}^{0i} - \mathrm{TP}^{0i}) & (\mathrm{FN}^{1i} - \mathrm{TN}^{1i}) & (\mathrm{FP}^{1i} - \mathrm{TP}^{1i}) \end{array} \right] \\ \mathbf{n}_i^T &= \frac{1}{p_i}\left[\begin{array}{cccc} (\mathrm{FN}^{0i} - \mathrm{TN}^{0i}) & (\mathrm{FP}^{0i} - \mathrm{TP}^{0i}) & (\mathrm{FN}^{1i} - \mathrm{TN}^{1i}) & (\mathrm{FP}^{1i} - \mathrm{TP}^{1i}) \end{array} \right] \\ \mathbf{z}_i^T &= \left[\begin{array}{cccc} z_0^{0i} & z_1^{0i} & z_0^{1i} & z_1^{1i} \end{array} \right] \\ \mathbf{m}_i^T &= \left[\begin{array}{cccc} \frac{-\mathrm{FN}^{0i}}{\alpha} & \frac{\mathrm{TP}^{0i}}{\alpha} & \frac{\mathrm{FN}^{1i}}{\beta} & \frac{-\mathrm{TP}^{1i}}{\beta} \end{array} \right] \\ b_i &= \frac{1}{p_i}(\mathrm{TN}^{0i} + \mathrm{TP}^{0i} + \mathrm{TN}^{1i} + \mathrm{TP}^{1i}) \end{aligned}$$

for $i = 1, 2, \cdots K$.

B. Theoretical Proof

B.1. Proof of Proposition 3.2

We first show that the outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}} : \mathcal{X} \times \mathcal{A} \times \mathcal{C} \to \mathcal{Y}$ satisfies equal opportunity:

The probability $\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=a)$ can be extended as:

$$
\begin{aligned}
&\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=a)\\
=&\frac{\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1,Y=1,A=a)}{\Pr_D(Y=1,A=a)}\\
=&\frac{\sum_{c=1}^{K}\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1,Y=1,A=a,C=c)}{\Pr_D(Y=1,A=a)}\\
=&\frac{\sum_{c=1}^{K}(\Pr_D(\widehat{Y}=1,Y=1,A=a,C=c)\cdot\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=\widehat{Y}|\widehat{Y}=1,A=a,C=c))}{\Pr_D(Y=1,A=a)}\\
+&\frac{\sum_{c=1}^{K}(\Pr_D(\widehat{Y}=0,Y=1,A=a,C=c)\cdot\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}\neq\widehat{Y}|\widehat{Y}=0,A=a,C=c))}{\Pr_D(Y=1,A=a)}\\
=&\frac{\sum_{c=1}^{K}(\Pr_D(\widehat{Y}=1,Y=1,A=a,C=c)\cdot z_1^{ac}}{\Pr_D(Y=1,A=a)}+\frac{\sum_{c=1}^{K}(\Pr_D(\widehat{Y}=0,Y=1,A=a,C=c)\cdot(1-z_0^{ac})}{\Pr_D(Y=1,A=a)}\\
=&\frac{\sum_{c=1}^{K}\mathrm{TP}^{ac}\cdot z_1^{ac}}{\Pr_D(Y=1,A=a)}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{ac}\cdot(1-z_0^{ac})}{\Pr_D(Y=1,A=a)}
\end{aligned}
\tag{16}
$$

We can now calculate the *Equal opportunity Difference* of the outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}}$, which is defined as: $\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=0)-\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=1)$:

$$
\begin{aligned}
&\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=0)-\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=1)\\
=&\frac{\sum_{c=1}^{K}\mathrm{TP}^{0c}\cdot z_1^{0c}}{\Pr_D(Y=1,A=0)}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{0c}\cdot(1-z_0^{0c})}{\Pr_D(Y=1,A=0)}-(\frac{\sum_{c=1}^{K}\mathrm{TP}^{1c}\cdot z_1^{1c}}{\Pr_D(Y=1,A=1)}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{1c}\cdot(1-z_0^{1c})}{\Pr_D(Y=1,A=1)})\\
=&\frac{-\sum_{c=1}^{K}\mathrm{FN}^{0c}\cdot z_0^{0c}}{\alpha}+\frac{\sum_{c=1}^{K}\mathrm{TP}^{0c}\cdot z_1^{0c}}{\alpha}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{1c}\cdot z_0^{1c}}{\beta}-\frac{\sum_{c=0}^{K}\mathrm{TP}^{1c}\cdot z_1^{1c}}{\beta}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{0c}}{\alpha}-\frac{\sum_{c=1}^{K}\mathrm{FN}^{1c}}{\beta}
\end{aligned}
\tag{17}
$$

The first linear equation of $\mathbf{Az}=\mathbf{b}$ in (8) is:

$$
\begin{aligned}
0=&\sum_{c=1}^{K}\mathbf{m}_c^T\mathbf{z}_c-\sum_{k=1}^{K}\left(\frac{\mathrm{FN}^{1k}}{\beta}-\frac{\mathrm{FN}^{0k}}{\alpha}\right)\\
=&\sum_{c=0}^{K}\begin{bmatrix}\frac{-\mathrm{FN}^{0c}}{\alpha} & \frac{\mathrm{TP}^{0c}}{\alpha} & \frac{\mathrm{FN}^{1c}}{\beta} & \frac{-\mathrm{TP}^{1c}}{\beta}\end{bmatrix}^T\cdot\begin{bmatrix}z_0^{0c}\\ z_1^{0c}\\ z_0^{1c}\\ z_1^{1c}\end{bmatrix}-\sum_{c=1}^{K}\left(\frac{\mathrm{FN}^{1c}}{\beta}-\frac{\mathrm{FN}^{0c}}{\alpha}\right)\\
=&\frac{-\sum_{c=1}^{K}\mathrm{FN}^{0c}\cdot z_0^{0c}}{\alpha}+\frac{\sum_{c=1}^{K}\mathrm{TP}^{0c}\cdot z_1^{0c}}{\alpha}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{1c}\cdot z_0^{1c}}{\beta}-\frac{\sum_{c=0}^{K}\mathrm{TP}^{1c}\cdot z_1^{1c}}{\beta}+\frac{\sum_{c=1}^{K}\mathrm{FN}^{0c}}{\alpha}-\frac{\sum_{c=1}^{K}\mathrm{FN}^{1c}}{\beta}
\end{aligned}
$$

Combine the above with the *Equal Opportunity Difference*'s expression (17) :

$$
\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=0)-\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}}=1|Y=1,A=1)=0
\tag{18}
$$

We can see from (18) the first linear equation leads a outcome predictor that satisfies equal opportunity.

Then, we show that the outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}}:\mathcal{X}\times\mathcal{A}\times\mathcal{C}\to\mathcal{Y}$ satisfies community fairness.

The condition for community fairness in Definition 2.3 is equivalent to:

$$\forall k \in \mathcal{C}, \Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = k) = \frac{1}{K} \sum_{c=1}^{K} \Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = c)$$

The error rate of community k: $\Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = k)$ can be extended as:

$$\begin{aligned}
\Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = k) &= \sum_{a=0}^{1} (\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = 0, Y = 1, A = a | C = k) + \Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = 1, Y = 0, A = a | C = k)) \\
&= \sum_{a=0}^{1} [\Pr_D(\widehat{Y} = 0, Y = 1, A = a | C = k) \cdot \Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = \widehat{Y} | \widehat{Y} = 0, A = a, C = k) \\
&+ \Pr_D(\widehat{Y} = 1, Y = 1, A = a | C = k) \cdot \Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq \widehat{Y} | \widehat{Y} = 1, A = a, C = k) \\
&+ \Pr_D(\widehat{Y} = 1, Y = 0, A = a | C = k) \cdot \Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = \widehat{Y} | \widehat{Y} = 1, A = a, C = k) \\
&+ \Pr_D(\widehat{Y} = 0, Y = 0, A = a | C = k) \cdot \Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq \widehat{Y} | \widehat{Y} = 0, A = a, C = k)] \\
&= \sum_{a=0}^{1} (\mathrm{FN}^{ak} \cdot z_0^{ak} + \mathrm{TP}^{ak} \cdot (1 - z_1^{ak}) + \mathrm{FP}^{ak} \cdot z_1^{ak} + \mathrm{TN}^{ak} \cdot (1 - z_0^{ak}))/\Pr(C = k) \\
&= \sum_{a=0}^{1} ((\mathrm{FN}^{ak} - \mathrm{TN}^{ak}) \cdot z_0^{ak} + (\mathrm{FP}^{ak} - \mathrm{TP}^{ak}) \cdot z_1^{ak} + (\mathrm{TP}^{ak} + \mathrm{TN}^{ak})) \cdot \frac{1}{p_k} \\
&= \mathbf{n}_k^T \cdot \mathbf{z}_k + b_k
\end{aligned} \tag{19}$$

The last n linear equations of $\mathbf{Az} = \mathbf{b}$ in (8) are:

for $k = 1, 2, 3, \cdots K$:

$$\begin{aligned}
0 &= -\frac{K-1}{K} \mathbf{n}_k^T \mathbf{z}_k + \frac{1}{K} \sum_{(c \in \mathcal{C}, c \neq k)} \mathbf{n}_c^T \mathbf{z}_c - \frac{1}{K} \sum_{c=1}^{K} (b_k - b_c) \\
&= -(\mathbf{n}_k^T \mathbf{z}_k + b_k) + \frac{1}{K} \sum_{c \in \mathcal{C}} (\mathbf{n}_c^T \mathbf{z}_c^T + b_c) \\
&= -\Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = k) + \frac{1}{K} \sum_{c=0}^{K} \Pr(\widetilde{Y}_{\widehat{Y},\mathbf{z}} \neq Y | C = c)
\end{aligned} \tag{20}$$

The last equation is from (19).

We can see from (20) the last K linear equations of $\mathbf{Az} = \mathbf{b}$, the outcome predictor satisfies community fairness.

From the proceeding, if $\mathbf{z} \in \mathbb{R}^{4K}$ is the solution of the linear program (8), the outcome predictor that satisfies (7) is a fair outcome predictor. The outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}}$ that takes values of (9) has:

$$\begin{aligned}
\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = \widehat{Y} | \widehat{Y} = 1, A = a, C = c) &= z_1^{ac} \\
\Pr_D(\widetilde{Y}_{\widehat{Y},\mathbf{z}} = \widehat{Y} | \widehat{Y} = 0, A = a, C = c) &= z_0^{ac}
\end{aligned}$$

which satisfies (9).

Thus, the outcome predictor $\widetilde{Y}_{\widehat{Y},\mathbf{z}}$ is a fair outcome predictor w.r.t. both equal opportunity and community fairness. $\diamondsuit$

B.2. Proof of Theorem 3.4

We show that the linear program (12) always has solutions. Before presenting the proof, we first present Farkas' lemma (Goldman & Tucker, 1957): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then exactly one of the following two assertions is true:

1. There exists a $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{z} = \mathbf{b}$ and $\mathbf{z} \geq 0$.

2. There exists a $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{A}^T\mathbf{y} \geq 0$ and $\mathbf{b}^T\mathbf{y} < 0$.

Faraks' lemma states that either the system $\mathbf{A}\mathbf{z} = \mathbf{b}$ has a non-negative solution or the system $\mathbf{A}^T\mathbf{y} \geq 0$ has a solution with $\mathbf{b}^T\mathbf{y} < 0$ but not both. Thus, we can show the linear program (12) always exist a solution by showing that the set $\{\mathbf{y} | \mathbf{y} \in \mathbb{R}^{5K+1}, \bar{\mathbf{A}}^T\mathbf{y} \geq 0, \bar{\mathbf{b}}^T\mathbf{y} < 0\}$ is always empty.

Let: $\mathbf{y}^T = \begin{bmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{bmatrix}$, with $\mathbf{y}_1 \in \mathbb{R}^{K+1}$, $\mathbf{y}_2 \in \mathbb{R}^{4K}$, then, the condition $\bar{\mathbf{A}}^T\mathbf{y} \geq 0$ is:

$$\begin{aligned} \bar{\mathbf{A}}^T\mathbf{y} &= \begin{bmatrix} \mathbf{A}^T & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}^T\mathbf{y_1} + \mathbf{I}\mathbf{y_2} \\ \mathbf{y_2} \end{bmatrix} \geq 0 \\ \rightarrow \quad & \mathbf{1}_{4K}^T \cdot (\mathbf{A}^T\mathbf{y_1} + \mathbf{y_2}) \geq 0 \\ & \mathbf{1}_{4K}^T \cdot \mathbf{y_2} \geq 0 \end{aligned} \tag{21}$$

The condition $\bar{\mathbf{b}}^T\mathbf{y} < 0$ is:

$$\begin{aligned} \bar{\mathbf{b}}^T\mathbf{y} &= \begin{bmatrix} \mathbf{b}^T & \mathbf{1}_{4K}^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} \\ &= \mathbf{b}^T\mathbf{y_1} + \mathbf{1}_{4K}^T\mathbf{y_2} \\ &= \frac{1}{2}\mathbf{A}^T\mathbf{y_1} + \mathbf{1}_{4K}^T\mathbf{y_2} < 0 \end{aligned} \tag{22}$$

The last equation above is from the fact: $\frac{1}{p_c}(\mathrm{TN}^{0c} + \mathrm{TP}^{0c} + \mathrm{TN}^{1c} + \mathrm{TP}^{1c}) = 1 - \frac{1}{p_c}(\mathrm{FN}^{0c} + \mathrm{FP}^{0c} + \mathrm{FN}^{1c} + \mathrm{FP}^{1c})$.

When $\mathbf{1}_{4K}^T\mathbf{y_2} \geq 0$, which is the second condition in (21), the first condition in (21): $\mathbf{1}_{4K}^T \cdot (\mathbf{A}^T\mathbf{y_1}) \geq -\mathbf{1}_{4K}^T\mathbf{y_2}$ is always conflict with the condition (22): $\mathbf{1}_{4K}^T \cdot (\mathbf{A}^T\mathbf{y_1}) < -2 \cdot \mathbf{1}_{4K}^T\mathbf{y_2}$, as $-2 \cdot \mathbf{1}_{4K}^T\mathbf{y_2} \leq -\mathbf{1}_{4K}^T\mathbf{y_2}$.

Thus, the set: $\{\mathbf{y} | \mathbf{y} \in \mathbb{R}^{5K+1}, \bar{\mathbf{A}}^T\mathbf{y} \geq 0, \bar{\mathbf{b}}^T\mathbf{y} < 0\}$ is always empty. This indicates the system $\bar{\mathbf{A}}\bar{\mathbf{z}} = \bar{\mathbf{b}}$ always has non-negative solutions. The variables in the linear program represent probabilities that are bounded in $[0, 1]$. Therefore, the objective function of (12) is bounded. The linear program always has solutions.

$\diamondsuit$

B.3. Proof of Theorem 3.5

Proof: The $\Pr_D(\widetilde{Y} \neq Y)$ in first condition can be extended as:

$$\begin{aligned}
\Pr_D(\widetilde{Y} \neq Y) &= \sum_{c=1}^{K}\sum_{a=0}^{1}(\Pr_D(\widetilde{Y}=0, Y=1, A=a, C=k) + \Pr_D(\widetilde{Y}=1, Y=0, A=a, C=k)) \\
&= \sum_{c=1}^{K}\sum_{a=0}^{1}[\Pr_D(\widehat{Y}=0, Y=1, A=a, C=k)\cdot \Pr_D(\widetilde{Y}=\widehat{Y}|\widehat{Y}=0, A=a, C=k) \\
&+ \Pr_D(\widehat{Y}=1, Y=1, A=a, C=k)\cdot \Pr_D(\widetilde{Y}\neq\widehat{Y}|\widehat{Y}=1, A=a, C=k) \\
&+ \Pr_D(\widehat{Y}=1, Y=0, A=a, C=k)\cdot \Pr_D(\widetilde{Y}=\widehat{Y}|\widehat{Y}=1, A=a, C=k) \\
&+ \Pr_D(\widehat{Y}=0, Y=0, A=a, C=k)\cdot \Pr_D(\widetilde{Y}\neq\widehat{Y}|\widehat{Y}=0, A=a, C=k)] \\
&= \sum_{c=1}^{K}\sum_{a=0}^{1}(\mathrm{FN}^{ak}\cdot z_0^{ak} + \mathrm{TP}^{ak}\cdot(1-z_1^{ak}) + \mathrm{FP}^{ak}\cdot z_1^{ak} + \mathrm{TN}^{ak}\cdot(1-z_0^{ak})) \\
&= \mathbf{c}^T\mathbf{z} + \sum_{c=0}^{K} b_c p_c \\
&= \bar{\mathbf{c}}^T\bar{\mathbf{z}} + \sum_{c=0}^{K} b_c p_c
\end{aligned} \tag{23}$$

As we show in proposition (3.2), the constraints of equal opportunity and community fairness in the linear program (8) (or a standard form (12)) are:

$$\begin{aligned}
&\bar{\mathbf{A}}\bar{\mathbf{z}} = \bar{\mathbf{b}} \\
\Rightarrow\quad &\bar{\mathbf{A}}^T\bar{\mathbf{A}}\bar{\mathbf{z}} = \bar{\mathbf{A}}^T\bar{\mathbf{b}}
\end{aligned}$$

For matrix $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$, we have:

$$\lambda_{\min}(\bar{\mathbf{A}}^T\bar{\mathbf{A}})\bar{\mathbf{z}}^T\bar{\mathbf{z}} \leq \bar{\mathbf{z}}^T\bar{\mathbf{A}}^T\bar{\mathbf{A}}\bar{\mathbf{z}} \tag{24}$$

where, $\lambda_{\min}(\bar{\mathbf{A}}^T\bar{\mathbf{A}})$ is the smallest eigenvalue of $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$.

Then, we show the l_2 norm of $\bar{\mathbf{z}}$ is bounded:

$$\begin{aligned}
&\lambda_{\min}(\bar{\mathbf{A}}^T\bar{\mathbf{A}})\bar{\mathbf{z}}^T\bar{\mathbf{z}} \leq \bar{\mathbf{z}}^T\bar{\mathbf{A}}^T\bar{\mathbf{A}}\bar{\mathbf{z}} \leq \|\bar{\mathbf{z}}\|_2\|\bar{\mathbf{A}}^T\bar{\mathbf{A}}\bar{\mathbf{z}}\|_2 = \|\bar{\mathbf{z}}\|_2\|\bar{\mathbf{A}}^T\bar{\mathbf{b}}\|_2 \\
\Rightarrow\quad &\lambda_{\min}(\bar{\mathbf{A}}^T\bar{\mathbf{A}})\|\bar{\mathbf{z}}\|_2 \leq \|\bar{\mathbf{A}}^T\bar{\mathbf{b}}\|_2 \\
\Rightarrow\quad &\|\mathbf{z}\|_2 \leq \frac{\|\bar{\mathbf{A}}^T\bar{\mathbf{b}}\|_2}{\underline{\sigma}^2}
\end{aligned} \tag{25}$$

where, $\underline{\sigma}$ is the smallest singular value of the matrix $\bar{\mathbf{A}}$.

Thus, the $\Pr_D(\widetilde{Y} \neq Y)$ has:

$$\begin{aligned}
\Pr_D(\widetilde{Y} \neq Y) &= \bar{\mathbf{c}}^T\bar{\mathbf{z}} + \sum_{c=0}^{K} b_c p_c \\
&\geq -\|\bar{\mathbf{c}}\|_\infty\|\bar{\mathbf{z}}\|_\infty + \sum_{c=0}^{K} b_c p_c \quad \textit{(all elements in c are negative)} \\
&\geq -\|\bar{\mathbf{c}}\|_\infty\|\bar{\mathbf{z}}\|_2 + \sum_{c=0}^{K} b_c p_c \quad (\|\bar{\mathbf{z}}\|_2 \geq \|\bar{\mathbf{z}}\|_\infty) \\
&\geq -\|\bar{\mathbf{c}}\|_\infty\frac{\|\bar{\mathbf{A}}^T\bar{\mathbf{b}}\|_2}{\underline{\sigma}^2} + \sum_{c=0}^{K} b_c p_c \quad \textit{(the upper bound (25))}
\end{aligned} \tag{26}$$

with $b_c p_c = (\mathrm{TN}^{0c} + \mathrm{TP}^{0c} + \mathrm{TN}^{1c} + \mathrm{TP}^{1c})$ as list in Appendix A.

When $\Delta < -\|\bar{\mathbf{c}}\|_\infty \frac{\|\bar{\mathbf{A}}^T \bar{\mathbf{b}}\|_2}{\underline{\sigma}^2} + \sum_{c=0}^{K} b_c p_c$, the inequality: $\Pr_D(\widetilde{Y} \neq Y) \geq -\|\bar{\mathbf{c}}\|_\infty \frac{\|\bar{\mathbf{A}}^T \bar{\mathbf{b}}\|_2}{\underline{\sigma}^2} + \sum_{c=0}^{K} b_c p_c$ is conflict with the first condition in (13): $\Pr_D(\widetilde{Y} \neq Y) \leq \Delta$.

Thus, if $\Delta < -\|\bar{\mathbf{c}}\|_\infty \frac{\|\bar{\mathbf{A}}^T \bar{\mathbf{b}}\|_2}{\underline{\sigma}^2} + \sum_{c=0}^{K} b_c p_c$, the fairness condition and ϵ-accurate condition are incompatible with each other. $\diamondsuit$

B.4. Proof of Theorem 3.6

The error rate of the fair outcome predictor is demonstrated in (23), which is: $\Pr_D(\widetilde{Y} \neq Y) = \mathbf{c}^T \mathbf{z} + \sum_{c=0}^{K} b_c \cdot p_c$.

The predictor $\widetilde{Y}$ is optimal when $\mathbf{z} = \mathbf{1}_{4k}$, the error rate of the optimal predictor is: $\Pr_D(\widehat{Y} \neq Y) = \mathbf{c}^T \mathbf{1}_{4K} + \sum_{c=0}^{K} b_c \cdot p_c$.

Thus, the minimum error we need to compromise for enforcing group fairness and community fairness is:

$\Pr_D(\widetilde{Y} \neq Y) - \Pr_D(\widehat{Y} \neq Y) = \mathbf{c}^T \mathbf{z} - \mathbf{c}^T \mathbf{1}_{4K}$. $\diamondsuit$

C. Additional Experimental Details and results

C.1. Additional Experimental Details

For UCI Adult, in each local community, we randomly divide the data into three subsets: 60% for the training set, 20% for the validation set, and 20% for the test set. We first implement the *FedAvg* algorithm. For each communication round in *FedAvg*, the number of participating communities is set to $N = 2$. We set the number of local update epochs to $E = 1$ with a batch size of $B = 128$. The local models are logistic regression classifiers with two layers, containing 64 and 32 nodes, respectively. We use *Relu* as the activation functions for each hidden layers. These models are trained using the Adam optimizer with a learning rate of $\eta = 0.05$. We select the number of rounds that minimize the disparity between training and validation accuracy, and then report the evaluation metrics on the test dataset. We construct a linear program using the training data. Finally, we apply post-processing to the test dataset based on the solution from the linear program and report its evaluation metrics.

For the Diabetes dataset, we similarly split the local dataset into 60% for training, 20% for validation, and 20% for testing. The number of participating communities for *FedAvg* is set to $N = 7$. We maintain the number of local update epochs at $E = 1$ with a batch size of $B = 256$. We follow the same model structure, optimization algorithm, and evaluation process as with the UCI Adult dataset and report its evaluation metrics.

We implement all code in TensorFlow (Abadi et al., 2015), simulating a federated network with one server and several local communities.

C.2. Baselines

- *q-FedAvg* (Li et al., 2019) improves community fairness in Federated Learning (FL) by minimizing an aggregated reweighted loss, parameterized by q. The algorithm assigns greater weight to devices with higher loss. The parameter q controls the trade-off between community fairness and model utility. In our experiments, we set $q = 4$, following the recommendation in the original implementation of the *q-FedAvg* paper.

- *FairFed*, (Ezzeldin et al., 2023) improves equal opportunity in FL, which also minimizes an aggregated reweighed loss, parameterized by β. The weights are a function of the mismatch between the global EOD (on the full dataset) and the local EOD at each community, favoring communities whose local measurement match the global measurements. β is the parameter that control the tradeoff between the group fairness and model utility. We follows the initial paper's setting: $\beta = 1$.

- *q-FedAvg+FairFed*: We build *q-FedAvg+FairFed*, a combination of *q-FedAvg* and *FairFed*. In the communication round t, the global model is set as $\boldsymbol{\theta}^t = \lambda \boldsymbol{\theta_1}^t + (1 - \lambda)\boldsymbol{\theta_2}^t$, with $\boldsymbol{\theta}_1^t$ representing the global model updated by *FairFed*, and $\boldsymbol{\theta}_2^t$ representing the global model updated by *q-FedAvg*.λ^t is the parameter that controls the balance between community fairness and group fairness. Setting $\lambda = 0$ recovers the *q-FedAvg* and setting $q = 1$ recovers the *FairFed*. We set λ to different values and report the result.

C.3. Convergence Curves

Figure 2. Adult dataset: Training curves of *Fedavg* (left), *FairFed* (middle), and *q-FedAvg* (right)