

Toward the Tradeoffs between Privacy, Fairness and Utility in Federated Learning

Kangkang Sun¹, Xiaojin Zhang², Xi Lin¹, Gaolei Li¹, Jing Wang¹, and Jianhua Li¹

Shanghai Key Laboratory of Integrated Administration Technologies for Information Security,
School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai, China

School of Computer Science and Technology, Huazhong University of Science and Technology,
Wuhan, China

{szpsunkk, linxi234, gaolei_li, wangjing08,
lijh888}@sjtu.edu.cn; xiaojinzhang@hust.edu.cn

Abstract. Federated Learning (FL) is a novel privacy-protection distributed machine learning paradigm that guarantees user privacy and prevents the risk of data leakage due to the advantage of the client's local training. Researchers have struggled to design fair FL systems that ensure fairness of results. However, the interplay between fairness and privacy has been less studied. Increasing the fairness of FL systems can have an impact on user privacy, while an increase in user privacy can affect fairness. In this work, on the client side, we use the fairness metrics, such as *Demographic Parity* (DemP), *Equalized Odds* (EOs), and *Disparate Impact* (DI), to construct the local fair model. To protect the privacy of the client model, we propose a privacy-protection fairness FL method. The results show that the accuracy of the fair model with privacy increases because privacy breaks the constraints of the fairness metrics. In our experiments, we conclude the relationship between privacy, fairness and utility, and there is a tradeoff between these.

Keywords: Fair and Private Federated Learning · Differential Privacy · Privacy Protection.

1 Introduction

Federated learning (FL) [MMR⁺17, KMA⁺21] is a novel distributed machine learning approach that guarantees user privacy by ensuring that user data does not leave the local area. However, FL has been plagued by two ethical issues: privacy and fairness [CZZ⁺23]. So far, most of the research has considered these two issues separately, but the existence of some kind of intrinsic equilibrium between the two remains unexplored. For example, privacy can come at the expense of model accuracy, however, for different groups of people training privacy results in different accuracies, with disadvantaged groups often suffering a greater cost in the training process. On the other hand, in order to ensure the fairness of the model and eliminate the bias in the training data or model [ABD⁺18, BHJ⁺21], the client needs to share more data with the server, which seriously increases the user privacy risk. Therefore, it is an open issue to investigate the

intrinsic connection between fairness and privacy in FL and to break the distress caused by its tradeoffs.

Privacy Destroys Fairness The first observation is that the decrease in accuracy due to deep DP models has a disproportionately negative impact on underrepresented subgroups [BPS19]. DP-SGD enhances model “bias” in different distributions that need to be learned. Subsequently, in the study [PMK⁺20], the impact of DP on fairness in three real-world tasks involving sensitive public data. There are significant differences in the model outputs when stronger privacy protections are implemented or when the population is small. Many works [TFVH21, EGLC22] have attempted to find reasons why privacy destroys fairness.

Fairness Increases Privacy Risk The client’s dataset is usually unbalanced and biased. This bias is gradually amplified during the machine learning process. For example, when a model is trained for accuracy, the model’s predictions will correlate with gender, age, skin, and race in a certain demographic group [ZVRG17, BHJ⁺21, Cho17].

Privacy and fairness are two important concepts in FL, and violating either one is unacceptable. Therefore, this paper explores the intrinsic relationship between privacy and fairness in FL and designs a privacy protection method for fair federated learning, to improve the model learning performance while ensuring the privacy and fairness constraint.

Relationship of fairness and privacy. We attempt to explore the relationship between fairness and privacy in FL. Intuitively, there is some intrinsic connection between fairness and privacy, and some balance between fairness, privacy, and utility.

- *Fairness:* We consider three fairness metrics, including Demographic Parity (DemP), Equalized Odds (EO) and Disparate Impact (DI). Comparing the research [PMK⁺20], we design the optimization function to be more complex, taking into account privacy and fairness constraints.
- *Privacy:* In this paper, we consider privacy-protection methods for fair Federated Learning based differential privacy.

Our contributions can be summarized as follows:

- A privacy-protection fairness FL method is proposed, in order to protect the model privacy of the client while sharing model parameters. Our proposed method is mainly divided into two parts: fairness training and privacy-protection training. Specifically, the client first trains a fairness proxy model and then trains a privacy-protection model based on that proxy model.
- In this paper, We experimentally obtained that the increase in privacy destroys the fairness of the model but appropriately increases the accuracy of the model. In order to improve the accuracy of the model and to ensure the fairness of the model, we designed private fair algorithms 2.
- We demonstrate the superiority of our proposed method and algorithms based on *Adult* datasets comparing popular benchmark *FedAvg* algorithms. Experiments prove that our algorithm can effectively guarantee model privacy in fair FL.

2 Related Work

2.1 Fairness of FL

Fairness of FL is defined in two ways: client fairness [LSBS19, MBS20, YLL⁺20, KKM⁺20] and algorithmic fairness [HPS16]. Algorithmic fairness has been extensively studied in traditional centralized machine learning through debiasing methods [KMA⁺21]. However, due to the fact that in FL, the server does not have access to client-side local data, it is already difficult to estimate the global data distribution simply by debiasing either server-side or client-side [MMR⁺17]. Much research has focused on client fairness in FL, such as in augmenting client data aspect [Hao21, JOK⁺18], in the client data distribution aspect [DLC⁺20, WKNL20]. From a model perspective, training a separate fairness model for each client is an open problem.

2.2 Privacy of FL

Many recent studies have focused on FL privacy risks [GMS⁺23, LGR23a, SLS⁺23, BWD⁺22]. A diversity of privacy-protection techniques have been proposed to discourage the risk of privacy leakage for users, including cryptographic techniques and the perturbation approach [CZZ⁺23]. Cryptographic approaches allow computation on encrypted data and provide strict privacy guarantees. However, they are computationally expensive compared to non-encryption methods [XBJ21]. This computational overhead seriously affects the machine learning training process, especially with a large number of parameters in the model. Therefore, the current state-of-the-art privacy-protection methods are perturbation-based, such as the DP mechanism [GKN17, WLD⁺20, WKL⁺21, SMS22]. The shuffler model is proposed to amplify the privacy of LDP's poor performance in comparison with the central DP mechanisms [RSL⁺08, EFM⁺19, CSU⁺19, BBGN20, GKG⁺21, GDD⁺21]. Most research based on Shuffler's model has focused on the study of tradeoffs between privacy, utility, and communication [CCKS22, GDD⁺21, LLF⁺23, ZXW⁺22, BBGN19]. However, there is very little research on the privacy protection of fair federated learning.

2.3 Fairness and Privacy of FL

Recently, some work [CZZ⁺23, PMK⁺20] has led to inconsistent reductions in accuracy due to private mechanisms for classification [FMST20] and generation tasks [GODC22]. Because of the tension between fairness and privacy, researchers often need to make trade-offs between the two perceptions [BPS19, EGLC22, TFVH21]. The trade-off may be to increase privacy preservation at the expense of fairness, i.e., by adopting a loose notion of fairness rather than a precise one or vice versa [BHJ⁺21, Cho17].

3 Preliminaries

3.1 Fairness in FL

We consider the following fairness metrics, including DemP, EO and DI. DemP denotes the same probability of getting a chance under some sensitive attribute. EO is a subset

Table 1: Private and Fair Federated Learning

References	Privacy Metrics	Fairness Metrics	Techniques		Trade-off type
			Privacy	Fairness	
[LZMV19]	ϵ -DP	EOs & DemP	Class conditional noise	Fairness constraints	I
[JKM ⁺ 19]	(ϵ, δ) -DP	EOs	Exponential mechanism & Laplace noise	Fairness constraints	/
[LGR23b]	(ϵ, δ) -DP	EOs & DemP	DP-SGDA	ERMI regularizer	II
[TFVH21]	(α, ϵ_p) -Renyi DP	EOs, AP & DemP	DP-SGD	Fairness constraints	II
[KGK ⁺ 18]	/	EA	MPC	Fairness constraints	II
[DGK ⁺ 22]	/	EOs	Proxy attribute	Post-processing	II
[WGN ⁺ 20]	/	DemP	Noisy attribute	Fairness constraints	II
[AKM20]	/	EOs	Noisy attribute	Post-processing	II
Our Method	(ϵ, δ) -DP	EOs, DemP, DI	Gaussian Noise	Fairness constraints	II

I: Trade fairness for privacy. II: Trade privacy for fairness.

EOs: Equalized Odds. DemP: Demographic Parity. AP: Accuracy Parity. EA: Equal Accuracy. DI: Disparate Impact.

of DP, defined as the probability of getting a chance on a given aspect is the same for different sensitive attributes. Let X, Y be the sensitive attribute and the true label, respectively. For example, $Y = 1$ often represents the condition of being able to apply for a loan, and $Y = 0$ is the condition of not meeting the loan. Thus, on the opportunity to apply for a loan, the output has the same probability for each person (characteristic), and then this is EO fairness.

Definition 1. (Demographic Parity (DemP)) [HPS16] We say that a predictor f satisfies demographic parity with respect to attribute A , instance space X and output space Y , if the output of the prediction $f(X)$ is independent of the sensitive attribute \mathcal{A} . For $\forall a \in A$ and $p \in \{0, 1\}$:

$$\mathbf{P}[f(X) = p \mid \mathcal{A} = a] = \mathbf{P}[f(X) = p] \quad (1)$$

Given $p \in \{0, 1\}$, for $\forall a \in A$:

$$\mathbb{E}[f(X) \mid \mathcal{A} = a] = \mathbb{E}[f(X)] \quad (2)$$

However, the left and right terms of the above equality are often not the same. Then, the loss l_{DemP} of DemP can be defined as follows:

$$l_{DemP} = \mathbb{E}[f(X) \mid \mathcal{A} = a] - \mathbb{E}[f(X)] \quad (3)$$

Definition 2. (Equalized Odds (EO)) [HPS16] We say that a predictor f satisfies equalized odds with respect to attribute A , instance space X and output space Y , if the output of the prediction $f(X)$ is independent of the sensitive attribute A with the label \mathcal{Y} . For $\forall a \in \mathcal{A}$ and $p \in \{0, 1\}$:

$$\mathbf{P}[f(X) = p \mid \mathcal{A} = a, Y = y] = \mathbf{P}[f(X) = p \mid Y = y] \quad (4)$$

Given $p \in \{0, 1\}$, for $\forall a \in A, y \in Y$:

$$\mathbb{E}[f(X) \mid \mathcal{A} = a, Y = y] = \mathbb{E}[f(X) \mid Y = y] \quad (5)$$

Then, the loss l_{EO} of EO can be defined as follows:

$$l_{EO} = \mathbb{E}[f(X) \mid \mathcal{A} = a, Y = y] - \mathbb{E}[f(X) \mid Y = y] \quad (6)$$

Remark 1. A binary predictor f , satisfying the demographic parity, is a special instance of equalized odds.

Definition 3. (Disparate Impact (DI)) [PMK⁺20] We say that a predictor f satisfies disparate impact with respect to attribute \mathcal{A} , if the output of the prediction $f(X)$ is independent of the sensitive attribute \mathcal{A} with a similar proportion of the different groups. For $a \in \{0, 1\}$, we have:

$$\min \left(\frac{\mathbf{P}(f(x) > 0 \mid a = 1)}{\mathbf{P}(f(x) > 0 \mid a = 0)}, \frac{\mathbf{P}(f(x) > 0 \mid a = 0)}{\mathbf{P}(f(x) > 0 \mid a = 1)} \right) = 1 \quad (7)$$

For $i \in [0, n]$ and i is a positive integer:

$$\min \left(\frac{\mathbf{P}(f(x) > 0 \mid a = i + 1)}{\mathbf{P}(f(x) > 0 \mid a = i)}, \frac{\mathbf{P}(f(x) > 0 \mid a = 0)}{\mathbf{P}(f(x) > 0 \mid a = n)} \right)_{i=0}^n = 1 \quad (8)$$

Then, the loss l_{DI} of DI can be defined as follows:

$$l_{DI} = \min \left(\frac{\mathbf{P}(f(x) > 0 \mid a = i + 1)}{\mathbf{P}(f(x) > 0 \mid a = i)}, \frac{\mathbf{P}(f(x) > 0 \mid a = 0)}{\mathbf{P}(f(x) > 0 \mid a = n)} \right)_{i=0}^n - 1 \quad (9)$$

3.2 Privacy in FL

The local dataset of clients contains sensitive data, which requires protecting the sensitive attributes while training. Differential Privacy (DP) is a privacy protection technique designed to safeguard individual data while allowing data analysis and mining [DR⁺14]. Local Differential Privacy (LDP) is deployed on clients to protect the attributes of the local dataset, in order to make sure that any algorithm built on this dataset is differentially private. The ϵ -differentially private mechanism \mathcal{M} is defined as follows:

Definition 4. (Local Differential Privacy (LDP)) [DR⁺14] A randomized algorithm $\mathcal{M} : X \rightarrow Y$ satisfies (ϵ, δ) -LDP with respect to a input set X and a noise output set Y , if $\forall x, x' \in X$ and $\forall y \in Y$ hold:

$$\mathbf{P}[\mathcal{M}(x) = y] \leq e^\epsilon \mathbf{P}[\mathcal{M}(x') = y] + \delta \quad (10)$$

Definition 5 (Gaussian Mechanism). Assume that a deterministic function $f : \mathcal{M}X \rightarrow Y$ with $\Delta_2(f)$ sensitivity, then for $\forall \delta \in (0, 1)$, random noise follows a normal distribution $\mathcal{N}(0, \sigma^2)$, the mechanism $\mathcal{M}(d) = f(d) + \mathcal{N}(0, \sigma^2)$ is (ϵ, δ) -DP, where

$$\epsilon \geq \frac{\sqrt{2 \ln(1.25/\delta)}}{\frac{\sigma}{\Delta_2 f}}, \quad \Delta_2(f) = \max_{d, d' \in \mathcal{D}} \|f(d) - f(d')\|_2 \quad (11)$$

3.3 Problem Formulation

There is a set of n clients in the FL system, where $m \in n$ clients are selected to participate in the FL training process. The clients have its own local dataset $\mathcal{D}_i = \{d_1, \dots, d_n\}$. Let $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$ denote the entire dataset and $f(\theta_i, d_i)$ as the loss function of client i , where the parameter $\theta \in \Theta$ is the model parameter. There are $m \in n$ clients. The clients are connected to an untrusted server in order to solve the ERM problem $F_i(\theta, \mathcal{D}_i) = \frac{1}{b} \sum_{j=1}^b f(\theta, d_{ij})$, where local estimated loss function dependent on the local dataset \mathcal{D}_i , and b is the local batch size. We give the ERM problem [KMA⁺21] in FL, as follows:

$$\begin{aligned} \arg \min_{\theta \in \mathcal{C}} \left(F(\theta) := \frac{1}{m} \sum_{i=1}^m F_i(\theta) \right), \\ \text{s.t. } l_{DemP} < \mu_{DemP}, \\ l_{EO} < \mu_{EO}, \\ l_{DI} < \mu_{DI}, \end{aligned} \quad (12)$$

where the l_{DemP}, l_{EO}, l_{DI} are the loss constraint of DemP, EO and DI, respectively. We use the Lagrangian multiplier [PMK⁺20] to transform the ERM problem (12) into a Min-Max problem, as follows:

$$\begin{aligned} F(\theta, \lambda, l) = \arg \min_{\theta_i \in \Theta} \max_{\lambda_{ij} \in \Lambda} \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{b} \sum_{j=1}^b f_i(\theta_i + d_{ij}) + \lambda_{ij} l_k \right\}, \\ k \in \{DemP, EO, DI\}, \end{aligned} \quad (13)$$

where the parameter $\lambda \in \Lambda$ is the Lagrangian multiplier. In this fairness stage, the purpose is to train the proxy model under the fairness matrixes, which is to solve the optimization problem. For the optimization problem (13), there is the Lagrangian duality between the following two functions:

$$\begin{aligned} \min_{\theta \in \Theta} \max_{\lambda \in \Lambda} F(\theta, \lambda, l), \\ \max_{\lambda \in \Lambda} \min_{\theta \in \Theta} F(\theta, \lambda, l). \end{aligned} \quad (14)$$

In order to solve the above dual optimization problem (14), many works assume the function F is Lipschitz and convex and obtain a ν -approximate saddle point of Lagrangian, with a pair $(\hat{\theta}, \hat{\lambda})$, where

$$\begin{aligned}
 F(\widehat{\theta}, \widehat{\lambda}, l) &\leq F(\theta, \widehat{\lambda}, l) + \nu \quad \text{for all } \theta \in \Theta, \\
 F(\widehat{\theta}, \widehat{\lambda}, l) &\geq F(\widehat{\theta}, \lambda, l) - \nu \quad \text{for all } \lambda \in \Lambda.
 \end{aligned}
 \tag{15}$$

Therefore we can get the Max-Min and the Min-Max dual problems are equivalent in the ERM problem (12). In order to search for the optimal value (θ^*, λ^*) (or *Nash Equilibrium* in-game) of the problem (12), many works study the fairness model by many approaches, such as the Zero-Game [JKM⁺19, MOS20], Distributionally Robust Optimization (DRO) [WGN⁺20], and Soft Group Assignments [WGN⁺20]. In this paper, the fair model is optimized by the DRO method through a Lagrangian dual multiplier in clients, and the model parameters are then transmitted to the server for model aggregation through privacy-protection.

4 Method

In this section, we design privacy protection for fair federated learning based on differential privacy. In section 4.1, the fair model in the FL system is obtained by the Algorithm 1, where the fair model of each client can be optimized under constraints of *DemP*, *EO* and *DI*. In section 4.2, we design a privacy protection algorithm 2 for the fair model optimized in section 4.1.

4.1 Fairness Predictor (Model) in Client

Firstly, the clients train their own personalized fairness predictor, and we designed an Algorithm 1 to train the fair model on each client. In the Algorithm 1 line 5 and line 7, the optimal values (θ^*, λ^*) are derived from the partial differential expression of the ERM problem (12). Secondly, each θ_i and λ_i update their own information according to the partial differential expression in Algorithm 1 line 6 and line 8. Finally, after time T_1 rounds, the fair model of the client i is output.

Algorithm 1 Fair-SGD for client

Input: Local loss function $f(\cdot)$, train dataset \mathcal{D}_i , learning rate η , batch size B

1: Initialize: $f_i(\theta_i) \leftarrow \text{random}$, $\lambda_i \leftarrow \text{max value}$

2: **for** Each client $i \in \mathcal{N}$ **do**

3: **for** $t \in T_1$ **do**

4: Take a random batch size B and $j \in B$

5: For θ_i : $\mathbf{g}_t(x_j) \leftarrow \nabla_{\theta_{(i,t)}} f_i(\cdot)$

6: $\theta_{(i,t+1)} \leftarrow \theta_{(i,t)} - \eta_t \mathbf{g}_t(x_j)$

7: For λ_i , $\mathbf{g}'_t(x_j) \leftarrow \nabla_{\lambda_{(i,t)}} f_i(\cdot)$

8: $\lambda_{(i,t+1)} \leftarrow \lambda_{(i,t)} + \eta g'_t(x_j)$

9: **end for**

10: **end for**

Output: Fair model $f_i(\theta_i)$

4.2 Privacy Protection Method in Fair FL

In this section, we design a privacy-protection fairness FL framework to protect the privacy and fairness of sensitive datasets in clients. As the above section, there is a trade-off between privacy, fairness and accuracy in the FL system. In this paper, we designed a privacy-protection algorithm, named FedLDP Algorithm 2, based on the FedAvg algorithm.

FedLDP: In the algorithm, we design to add differential privacy preservation to the fairness model training process in algorithm 2. The algorithm, while reducing privacy consumption, can effectively improve the training accuracy of the model. Moreover, the algorithm does not guarantee that the intermediate entities are trustworthy, so the shuffler model is hijacked or attacked without any impact on user privacy.

Algorithm 2 FedLDP

Input: Selected clients m , the local dataset \mathcal{D}_i of client i , Maximum L_2 norm bound C , local privacy budget ε_l

- 1: Initial the local model and download the global gradients from the server
- 2: **for** $i \in m$ in parallel **do**
- 3: Fairness stage in Algorithm (1)
- 4: $\mathbf{g}_t(x_j) \leftarrow \nabla_{\theta_{(i,t)}} f_i(\cdot)$
- 5: $\bar{\mathbf{g}}_t(x_j) \leftarrow \mathbf{g}_t(x_j) / \max\left(1, \frac{\|\mathbf{g}_t(x_j)\|_2}{C}\right)$
- 6: $\tilde{\mathbf{g}}_t(x_j) \leftarrow \frac{1}{B} (\sum_i \bar{\mathbf{g}}_t(x_j) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$
- 7: $\theta_{(i,t+1)} \leftarrow \theta_{(i,t)} - \eta_t \tilde{\mathbf{g}}_t(x_j)$
- 8: **end for**
- 9: **Server**
- 10: **Aggregate:** $\bar{\mathbf{g}}_t \leftarrow \frac{1}{N_t} \sum_{i \in \mathcal{N}_t} w_t(d_{ij})$
- 11: **Gradient Descent:** $\theta_{t+1}^G \leftarrow \theta_t^G + \bar{\mathbf{g}}_t$

5 Experiments

5.1 Dataset and Experimental Settings

In order to test the performance proposed in this paper, we use the *Adult* [PG20], which is extracted from the U.S. Census dataset database, which contains 48,842 records, with 23.93% of the annual income greater than \$50k and 76.07% of the annual income less than \$50k, and has been divided into 32,561 training data and 16,281 test data. The class variable of this dataset is whether the annual income is more than \$50k or not, and the attribute variables include 14 categories of important information such as age, type of work, education, occupation, etc., of which 8 categories belong to the category discrete variables and the other 6 categories belong to the numerical continuous variables. This dataset is a categorical dataset that is used to predict whether or not annual income exceeds \$50k. We choose race as the sensitive attribute, including white person and black person.

5.2 Experimental Hyperparameter Settings

In the experiment, each client applied three (100×100) fully connected layers.

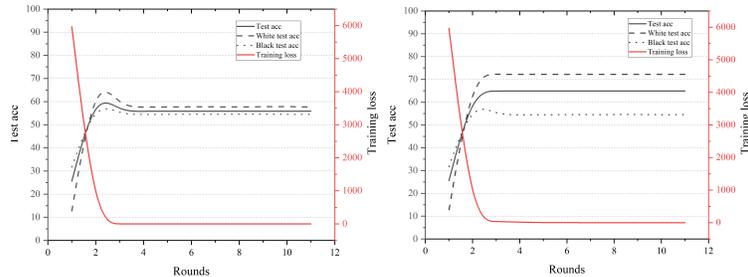
Machines The experiment was run on an ubuntu 2022.04 system with an intel i9 12900K CPU, GeForce RTX 3090 Ti GPU, and pytorch 1.12.0, torchvision 0.13.0, python 3.8.13.

Software We implement all code in PyTorch and the fair_learn tool.

5.3 Performance Comparison Results

In the experiment, we compared the test accuracy between different algorithms. In the FL system, we tested both cases of fairness training without noise, and fairness training with noise, shown in Fig. 1 (a) and (b). In Fig. 1 (a), the test accuracy of the white person is the same as the black person without noise in the client training process, while the fair client model with noise increases discrimination against different races in Fig. 1 (b).

Table 2 and Table 3 represent the test accuracy of differential clients in the FL system without noise and with noise, respectively. It can see from the table, that adding privacy improves the test accuracy for clients. The increase in privacy affects fairness because the increase in noise facilitates the optimizer to solve the global objective optimum while weakening the limitations of the fairness metrics, i.e., the constraints function $\lambda_{ij}l_k$.



(a) Fairness predictor with no privacy (b) Fairness predictor with privacy ($\mathcal{N}(0, 1)$)

Fig. 1: The average test accuracy of the fair stage training process in FL settings with 5 clients on *Adult* dataset. (a) and (b) are the training results with no privacy and privacy ($\mathcal{N}(0, 1)$), respectively. (a) is shown that the test accuracy of sensitive data *black* and *white* are approximately the same for both. With the addition of noise privacy, test accuracy improves but fairness decreases, shown in (b).

	Client 1	Client 2	Client 3	Client 4	Client 5
Black	32.20 %	69.42 %	68.80%	68.96%	33.36 %
White	12.26%	88.39 %	87.20%	87.05%	13.85 %

Table 2: The fair stage training process in FL settings with 5 clients (no privacy) on *Adult* dataset.

	Client 1	Client 2	Client 3	Client 4	Client 5
Black	66.63 %	73.75 %	68.96 %	69.41 %	67.79 %
White	86.11 %	85.70 %	87.05%	88.39%	87.73 %

Table 3: The fair stage training process in FL settings with 5 clients (privacy $\mathcal{N}(0, 1)$) on *Adult* dataset.

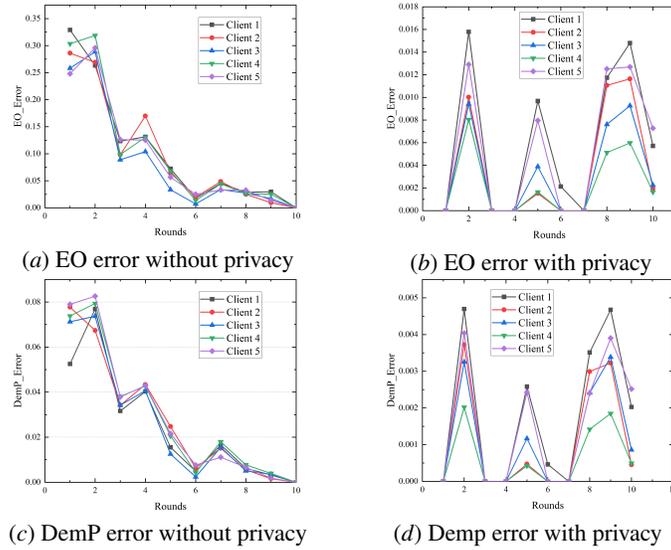


Fig. 2: The EO and DemP error comparison of different clients with privacy and no privacy on *Adult* dataset

5.4 Analysis of Privacy and Fairness

In this section, we analyse the influence of privacy and fairness on the client model. We analyse the fairness metrics of *EO Error* and *DemP Error* to evaluate the error of the training fairness model by adding the privacy ($\sigma = 1$). Fig. 2 (a)-(d) show the *EO Error* and *DemP Error* of different algorithms when each client trains the local fairness model and adds privacy noise. From Fig. 2 (a) and (c), the *EO Error* and *DemP Error* without privacy converge to zero. It can be shown that the client-trained model is fair in both *Demographic Parity* and *Equalized Odds*. However, when privacy is added during federated learning training, the *EO* l_{EO} and *DemP* l_{DemP} loss of the model does not

converge, which indicates that adding privacy to the model training process affects the fairness of the model.

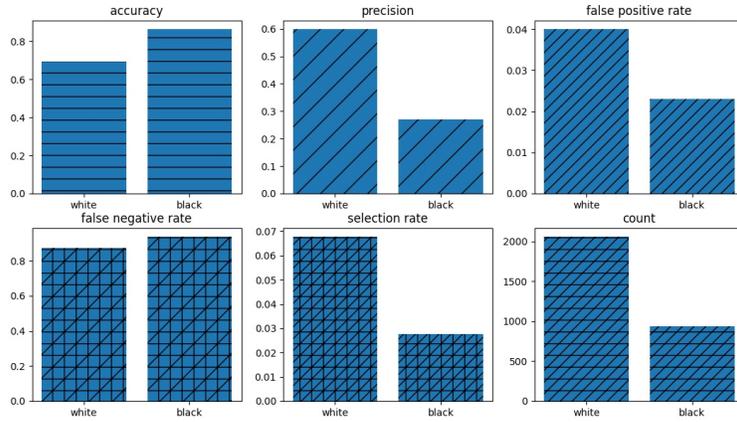
In Fig. 3, it is shown the fairness metrics in the client model with privacy and without privacy. In particular, client-side prediction performance is significantly increased by adding noise to the accuracy metric. One of the reasons for this is probably because, with the addition of privacy, the optimizer can jump out of the local optimum in finding the optimal solution.

6 Conclusion

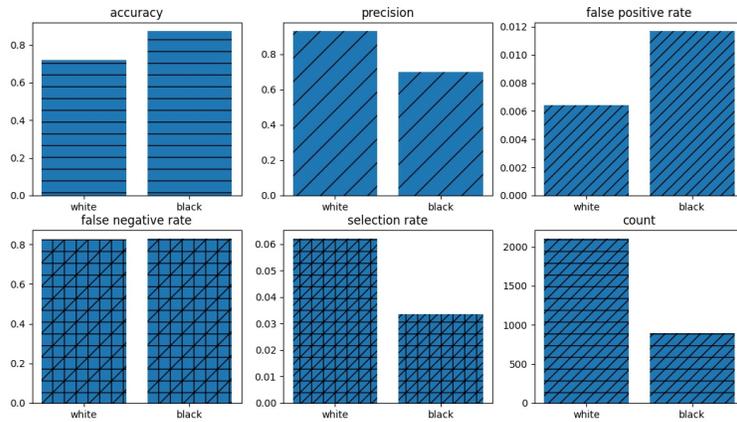
In this paper, we research the relationship between fairness and privacy in the FL system. Through the experiment, we found that there is a trade-off between privacy, fairness and accuracy in the FL system. In this paper, we construct the fairness model in clients under the fair metrics constraints, such as *Demographic Parity* (DemP) and *Equqlized Odds* (EOs). In order to protect the fair model privacy, we design a privacy-protecting fairness FL method and we give a private fair algorithm *FedLDP*. In our experiments, we conclude that by adding privacy we can appropriately increase the accuracy of the model while at the same time destroying its fairness.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant U20B2048, 62202302.



(a) Fair model of client 1 without privacy



(b) Fair model of client 1 with privacy ($\mathcal{N}(0, 1)$)

Fig. 3: The fairness metrics of clients on *Adult* dataset

Bibliography

- [ABD⁺18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.
- [AKM20] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics*, pages 1770–1780. PMLR, 2020.
- [BBGN19] Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Advances in Cryptology—CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II 39*, pages 638–667. Springer, 2019.
- [BBGN20] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 657–676, 2020.
- [BHJ⁺21] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.
- [BPS19] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.
- [BWD⁺22] Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, and Steven Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In *International Conference on Machine Learning*, pages 1945–1962. PMLR, 2022.
- [CCKS22] Wei-Ning Chen, Christopher A Choquette Choo, Peter Kairouz, and Ananda Theertha Suresh. The fundamental price of secure aggregation in differentially private federated learning. In *International Conference on Machine Learning*, pages 3056–3089. PMLR, 2022.
- [Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- [CSU⁺19] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology—EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pages 375–403. Springer, 2019.
- [CZZ⁺23] Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. Privacy and fairness in federated learning: on the perspective of trade-off. *ACM Computing Surveys*, 2023.

- [DGK⁺22] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1207–1239, 2022.
- [DLC⁺20] Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.
- [DR⁺14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [EFM⁺19] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.
- [EGLC22] Maria S Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C Cresswell. Disparate impact in differential privacy from gradient misalignment. *arXiv preprint arXiv:2206.07737*, 2022.
- [FMST20] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pages 15–19, 2020.
- [GDD⁺21] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021.
- [GGK⁺21] Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 463–488. Springer, 2021.
- [GKN17] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.
- [GMS⁺23] Till Gehlhar, Felix Marx, Thomas Schneider, Ajith Suresh, Tobias Wehrle, and Hossein Yalame. Safefl: Mpc-friendly framework for private and robust federated learning. *Cryptology ePrint Archive*, 2023.
- [GODC22] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pages 6944–6959. PMLR, 2022.
- [Hao21] Hao, Weituo and El-Khamy, Mostafa and Lee, Jungwon and Zhang, Jianyi and Liang, Kevin J and Chen, Changyou and Duke, Lawrence Carin. Towards fair federated learning with zero-shot data augmentation. In *Pro-*

- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3310–3319, 2021.
- [HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [JKM⁺19] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.
- [JOK⁺18] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.
- [K GK⁺18] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.
- [KKM⁺20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.
- [KMA⁺21] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- [LGR23a] Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pages 5749–5786. PMLR, 2023.
- [LGR23b] Andrew Lowy, Devansh Gupta, and Meisam Razaviyayn. Stochastic differentially private and fair learning. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, pages 86–119. PMLR, 2023.
- [LLF⁺23] Xiaochen Li, Weiran Liu, Hanwen Feng, Kunzhe Huang, Yuke Hu, Jinfei Liu, Kui Ren, and Zhan Qin. Privacy enhancement via dummy points in the shuffle model. *IEEE Transactions on Dependable and Secure Computing*, 2023.
- [LSBS19] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.
- [LZMV19] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. *Advances in neural information processing systems*, 32, 2019.
- [MBS20] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

- [MMR⁺17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [MOS20] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.
- [PG20] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.
- [PMK⁺20] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.
- [RSL⁺08] Sofya Raskhodnikova, Adam Smith, Homin K Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. What can we learn privately. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 531–540, 2008.
- [SLS⁺23] Jiawei Shao, Zijian Li, Wenqiang Sun, Tailin Zhou, Yuchang Sun, Lumin Liu, Zehong Lin, and Jun Zhang. A survey of what to share in federated learning: Perspectives on model utility, privacy leakage, and communication efficiency. *arXiv preprint arXiv:2307.10655*, 2023.
- [SMS22] Daniel Scheliga, Patrick Mäder, and Marco Seeland. Precode-a generic model extension to prevent deep gradient leakage. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1849–1858, 2022.
- [TFVH21] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021.
- [WGN⁺20] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33:5190–5203, 2020.
- [WKL⁺21] Yuezhou Wu, Yan Kang, Jiahuan Luo, Yuanqin He, and Qiang Yang. Fedcg: Leverage conditional gan for protecting privacy and maintaining competitive performance in federated learning. *arXiv preprint arXiv:2111.08211*, 2021.
- [WKNL20] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1698–1707. IEEE, 2020.
- [WLD⁺20] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learn-

- ing with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.
- [XBJ21] Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*, 2021.
- [YLL⁺20] Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 393–399, 2020.
- [ZVRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.
- [ZXW⁺22] Zan Zhou, Changqiao Xu, Mingze Wang, Xiaohui Kuang, Yirong Zhuang, and Shui Yu. A multi-shuffler framework to establish mutual confidence for secure federated learning. *IEEE Transactions on Dependable and Secure Computing*, 2022.