Information Sciences 619 (2023) 968–986

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Heterogeneity-aware fair federated learning

Xiaoli Li [a,b], Siran Zhao [a], Chuan Chen [a,*], Zibin Zheng [a]

[a] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[b] Computer School of HuBei University of Arts and Science, Xiangyang, China

ARTICLE INFO

Article history:
Received 6 May 2022
Received in revised form 15 September 2022
Accepted 12 November 2022
Available online 17 November 2022

Keywords:
Federated learning
Data heterogeneity
Fairness

ABSTRACT

The fairness of federated learning means that the global model cannot discriminate against any group. Due data heterogeneity, the update direction of some clients will hinder other clients, so the global model is difficult to treat each user fairly. The current fair federated learning methods usually use the variance of model performance to measure the fairness, which can not quantify the fairness of federated learning process; And, they all intervene from the beginning, which will reduce the convergence speed and model performance; Moreover, the weights they applied to clients are designed based on experience, so it is difficult to trade-off between model performance and fairness. To address the above problems, firstly, we use Gini coefficient to quantify the fairness of federated learning, which can reflect the fairness of federated learning process in each round. Secondly, we divide federated learning into two different training stages: label fitting and data fitting, and propose fairness intervention in the data fitting stage. Thirdly, we propose a fairness federated optimization algorithm, which introduces fairness penalty term into the objective function, and obtains the relationship between clients' gradients and fairness through gradient descent. Experimental results show the effectiveness and fairness of the proposed method.

1. Introduction

With the rapid growth of modern mobile and IoT (Internet of things) devices, a large amount of data has been generated. Machine learning based on these data has been applied to various walks of our lives. Conventional machine learning usually needs to fuse the data between devices on a central server for centralized training. However, as more and more users realize the problems of data security and user privacy, data integration faces great resistance. In recent years, Google has proposed federated learning [1,2], which can build a shared global machine learning model among a large number of distributed devices while keeping users' data locally, so it has gained increasingly extensive attention. Up to now, the most widely used method in federated learning is Federated Averaging algorithm (FedAvg) [3]. FedAvg mainly follows three steps in each communication iteration: (a) the server sends the latest global model to the client; (b) the clients receive the global model, train based on their local datasets, and then send the gradient updates to the server; (c) the server collects the clients' gradient updates to aggregate into a new global model. These steps are repeated until convergence.

In federated learning, as the data of the clients are generated independently, they usually show different distribution characteristics, that is, Non-Independent and Identically Distributed (Non-IID). As the updates are calculated by clients based on their own local datasets that are heterogeneous, there are divergence between client updates [4]. Therefore, the global

* Corresponding author.
E-mail address: chenchuan@mail.sysu.edu.cn (C. Chen).

https://doi.org/10.1016/j.ins.2022.11.031
0020-0255/

model that performs well on some clients may not perform well on some of the other clients, giving rise to severe fairness issue.

An intuitive way to alleviate this problem is to apply a relatively large weight to the updates of clients with poor effect. In the federated learning process, adjusting the weights of the clients will change the update direction of the global model, thereby indirectly affecting the loss of the global model. There are rich research lines on using the methods of weighting clients to solve the fairness problem. Huang et al. [5] proposed FedFa, which gives higher weights to the clients with lower training accuracy to encourage a more fair distribution of model performance across clients. Li et al. [6] proposed q-Fair Federated Learning (q-FFL), which thinks of the global model as a resource, and revises the objective function as α-fairness metric [7,8]. In such a setting, the clients with larger losses are given a relatively high weight, thus reducing the model performance difference of clients and inducing fairness. Li et al. [9] proposed tilted empirical risk minimization (TERM), which improves the worst-performing losses by paying a penalty on average performance, thus promoting a notion of uniformity or fairness. Although these works can alleviate the problem of unfairness in federated learning to a certain extent, they have the following problems: (a) they generally use the variance of the performances to the fairness of federated learning, which is an absolute indicator and is related to the accuracy size. The accuracy of federated learning changes with the training iterations, thus the variance of the performances can not quantitatively evaluate the fairness of federated learning timely; (b) they intervene in the process of federated learning from the beginning, that is, they weight the updates of the clients in the whole process of federated learning. However, as Geyer [10] presented, in the early stage of federated learning, the fluctuation between the updates of two iterations is relatively large, and the updates by clients are similar. Excessive fairness intervention at this stage not only has little effect on improving fairness, but also may affect the convergence speed. (c) it is difficult for them to make a trade-off between performance and fairness. The clients' weights in these methods are designed empirically. Generally, if they want to achieve greater fairness, they need to increase the weight difference between devices with good performance and devices with poor performance. However, sometimes when the weight difference is too large, it will significantly reduce the performance of the global model, and even make the global model unable to converge.

To address the above question, in this work, we define the fairness of federated learning as the Gini coefficient of the testing accuracy based on clients. This idea comes from the concept of "economic inequalities". The model performance of each client can be regarded as the income of the client, then we need to reduce the inequality between the income of each client. Furthermore, we divide the process of federated learning into two different training stages: label fitting and data fitting as Geyer [10] suggested, and propose to intervene the fairness in the data fitting stage. Finally, we propose a fairness federated optimization algorithm, which adds a fairness penalty term to the objective function, and obtains the relationship between clients' gradient updates and the fairness of the global model through a gradient descent algorithm.

The contributions of this paper are summarized as follows:

1. We adopt the Gini coefficient of the testing accuracy based on clients to quantify the fairness of federated learning. Gini coefficient is a relative indicatrix and strongly Lorenz-Consistent, which can report the amount of fairness of each communication round. That can not only help us understand the changes of fairness in the federated learning process, so as to adjust the weights of clients to achieve the expected fairness and model performance, but also better motivate clients to participate in the training process.
2. Different from the previous methods to intervene in the whole process of federated learning, we divide federated learning into two different training stages: label fitting and data fitting, and propose fairness intervention should be executed in the data fitting stage. Because in this stage, the updates fluctuation drastically shrinks, and the individual updates look less alike.
3. We propose a fairness federated optimization algorithm. Different from the previous empirical methods, we add a fairness penalty term to the objective function, and obtain the relationship between the clients' gradient updates and the global model fairness through gradient descent, so as to improve the fairness of the model while maintaining the performance of the global model.
4. To evaluate the advantages of our framework, we conduct experiments on some real datasets, and compare our approach with many state-of-the-art methods. The experimental results demonstrate the effectiveness of the proposed approach.

The remaining of this paper is organized as follows. Section 2 surveys related work. Section 3 introduces our method. Experimental results and analysis are summarized in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

Fairness in machine learning has been extensively studied. There are different types of fairness definitions. For example, Demographic parity [11] refers that for individuals in two different groups, the proportion assigned to each category should be the same; Individual fairness [11] means that similar individuals should have similar predictions; Equal opportunity [12] states that the true positive rates of different groups should be equal; Disparate mistreatment [13] states that refers the misclassification rates for groups having different values of sensitive attributes should be similar. These existing research on fairness in machine learning mostly focuses on protecting sensitive attributes of certain individuals or groups of individuals.

In federated learning, the fairness means treating federated learning participants fairly. Wang et al. [14] proposed that when there is a large divergence between conflicting updates, the global model is biased in favour of the client with a larger

update, resulting in a significant reduction in the model accuracy of the client with a smaller update. Mohri et al. [15] proposed that the target distribution of federated learning is not necessarily the uniform distribution of all clients' data. For example, the data of client is constantly generated. The client with a small amount of data during learning may produce a large amount of data later, which will affect the global data distribution. The conventional federation learning is biased towards the clients with a large amount of data or data distribution similar to the global data distribution, which is unfair. We summarize the fairness of federated learning into two categories: collaborative fairness and min–max fairness.

Collaborative fairness was proposed by Lyu et al. [16], in which different models are distributed to participant clients according to their contributions, and the high contributors should get better models than the low contributors. Xu et al. [17] proposed that some participant clients will be at a disadvantage in the contribution assessmen due to the Non-IID data. Zhang et al. [18] used a task-dependent strategy to measure contributions, in which the agents classify contributions according to some publicly verifiable factors, such as data quality, data volume, data collection cost, etc.

Min–max fairness in federated learning was proposed by Li et al. [6], it means that more equitable accuracy allocation between different clients. As Non-IID data in federated learning, min–max fairness is not to optimize all clients to obtain the same accuracy, but to reduce the difference of model accuracy of each client as much as possible. Li et al. [6] proposed q-FFL to give the clients with large errors a relatively high weight via an aggregate reweighted loss parameterized by q. However, in q-FFL, the accuracy may reduce due to Non-IID data; and the convergence will be unstable as the loss function has been expanded; moreover, using variance of the performance as the definition of fairness is not good enough, as it is an absolute indicator and is related to the accuracy size. Huang et al. [5] designed a weighting strategy based on the frequency of training participation and the training accuracy to encourage a more fair distribution of model performance across clients. Mohri et al. [15] proposed Agnostic Federated Learning (AFL), which optimizes the loss of the worst target distribution formed by a mixture of the client distributions. Du et al. [19] used kernel reweighting functions to assign a reweighting value to each training sample to ensure high accuracy and fairness for unknown test data.

Most of the current methods for the fairness of the global model are based on appropriate weight, which give a larger weight to the clients with higher losses. These weight-based methods need sufficient experience to adjust the weights, and the performances of the global model depends on how much weight is assigned to each client's local model. Our global model is gradient descent-based, and it can automatically tune the parameters of local models without a priori formula, thus alleviating the need to manually adjust the weights.

3. Method

In this section, we first introduce the classical federated learning method, and formally define the fairness of federated learning. Then, we propose an appropriate time to intervene the fairness. Finally, we introduce our fairness federated optimization algorithm.

3.1. Classical Federated Learning

The current most widely recognized method in federated learning is Federated Averaging algorithm (FedAvg) [3], which mostly follow three steps: (i) in each communication iteration, the server selects a random fraction of clients, and sends them the latest global model; (ii) the selected clients perform training based on their local data to update their local models, and then send their latest local client models to the server; (iii) the server collects a certain number of the local client models to aggreate a new global model. The steps repeate until convergence.

The goal of FedAvg is to minimize the empirical risk as follows:

$$F(\mathbf{w}) = \frac{1}{\sum_{k=1}^{K}|D_k|}\sum_{k=1}^{K}\sum_{\xi\in D_k} f(\mathbf{w},\xi), \tag{1}$$

where $\mathbf{w}$ is the global model parameter vector, K is the number of clients, D_k is the local dataset of client k, $|D_k|$ represents the number of samples in D_k, and $f(\mathbf{w},\xi)$ is the loss function of the global model parameter vector $\mathbf{w}$ in sample ξ. For client k, if we denote the fraction of local data $|D_k|/\sum_{k=1}^{K}|D_k|$ by p_k, and the local objective function $\sum_{\xi\in D_k} f(\mathbf{w},\xi)/|D_k|$ by $F_k(\mathbf{w})$, Eq. 1 can be rewritten as follows:

$$F(\mathbf{w}) = \sum_{k=1}^{K} p_k F_k(\mathbf{w}). \tag{2}$$

We assume that there are N clients $C_1,\ldots,C_N$ participating in the training in t-th round. Each client k calculates the local updates of the loss function with respect to the global model $\mathbf{w}$:

$$\Delta\mathbf{w}_k = \frac{\partial F_k(\mathbf{w})}{\partial \mathbf{w}}. \tag{3}$$

Then local updates $\Delta \mathbf{w}_k$ will be sent to the server. Then, the server aggregates the received local updates to refine the global model $\mathbf{w}$. The formulas for calculation are as follows:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \alpha_t \frac{1}{N} \sum_{k=1}^{N} \Delta \mathbf{w}_k, \tag{4}$$

where α_t is the learning rate.

3.2. Fairness of Federated Learning

Before introducing Gini coefficient, it is necessary to introduce Lorenz's Curve to help to understand Gini coefficient. Lorenz's Curve is a graph method to express income inequality. In this paper, the loss of each client is viewed as an income for that client. In federated learning, the loss inequality over multiple clients is minimized to make the global model more fair. In federated learning, clients are sorted from low to high according to their loss values. The horizontal axis of the graph is the cumulative value of client proportion, and the vertical axis of the graph is the cumulative value of loss proportion (from 0 to 1). Then the corresponding cumulative values are calculated respectively. Finally, the Lorenz's Curve can be obtained by making a scatter diagram and connecting these points with a smooth curve (see Fig. 1).

When the income distribution is absolutely equal, the two cumulative values are always equal, and the Lorentz curve is a straight line with a slope of 45° through the origin; when the income distribution is uneven, the Lorentz curve is below the straight line. Obviously, the closer the Lorentz curve is to the 45° straight line, the lower the degree of inequality; the farther away the Lorentz curve is from the straight line, the higher the degree of inequality. In the most extreme case, the curve coincides with the X and Y axes (moving the vertical axis to the right of the graph), and all income is obtained by the last client.

Gini Coefficient [20] refers to the common index used internationally to measure the income gap of residents in a country or region. The Gini coefficient can be calculated from the Lorenz's Curve in Fig. 1, which is equal to the area of part A in the figure divided by the area of part below the 45° line (i.e. A + B). Since there are only a limited number of clients participating in the training in federated learning, Gini coefficient can be defined as follows:

$$\begin{aligned} G(l) &= \frac{A}{A+B} = \frac{A+B-B}{A+B} = 1 - 2B \\ &= 1 - \frac{1}{K} \sum_{i=1}^{K} \left(\frac{\sum_{j=1}^{i-1} l_j + \sum_{j=1}^{i} l_j}{\sum_{j=1}^{K} l_j} \right) \end{aligned} \tag{5}$$

where K is the total number of clients, and the clients are sorted from low to high according to their loss, the loss of the client j is represented by l_j. The maximum Gini coefficient is 1 and the minimum is 0. The closer Gini coefficient is to 0, the more equal income distribution is.

Definition 1. ϵ-Fairness. We quantify the fairness of federated learning as the Gini coefficient of the testing accuracy for all clients:

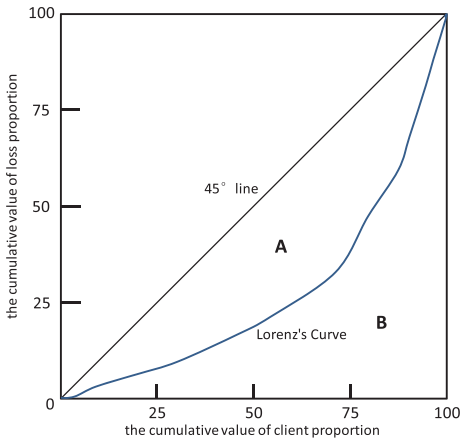


Fig. 1. The Lorenz's Curve in federated learning: The horizontal axis of the graph is the cumulative value of client proportion, and the vertical axis of the graph is the cumulative value of test accuracy proportion..

$$G(a_1,\ldots,a_K)=1-\frac{1}{K}\sum_{i=1}^{K}\left(\frac{\sum_{j=1}^{i-1}a_j+\sum_{j=1}^{i}a_j}{\bar{a}}\right). \tag{6}$$

In Eq. 6, the testing accuracy of the client j is represented by $a_j, \bar{a}=\sum_{i=1}^{K}a_i$ is the average of the testing accuracy for all client, and the clients are sorted from low to high according to testing accuracy, that is, $a_1<a_2<\ldots<a_K$. The maximum Gini coefficient is 1 and the minimum is 0. The closer Gini coefficient is to 0, the more fairness the federated learning is. Let $0\leqslant\epsilon\leqslant 1$, for a trained global model $\mathbf{w}$, if the Gini coefficient of the testing accuracy for all clients is equal to ϵ, we say that the global model $\mathbf{w}$ is ϵ-Fairness. For two trained models $\mathbf{w}$ and $\tilde{\mathbf{w}}$, and their fairness is ϵ and $\tilde{\epsilon}$ respectively, we say that the model $\mathbf{w}$ provides a more fair solution to $\tilde{\mathbf{w}}$ if $\epsilon<\tilde{\epsilon}$.

Gini coefficient is a relative indicatrix and strongly Lorenz-Consistent, while the variance of the testing accuray, which is proposed to quantify the fairness of federated learning by q-FFL [6], is an absolute indicator and is related to the accuracy size. For example, Table 1 shows the accuracy of the global model on 10 clients in communication round i,j,k, where $i>j>k$. The accuracy of clients $1\sim 10$ in $Round_j$ is $19\%\sim 10\%$ higher than that in $Round_i$. Similarly, the accuracy of clients $1\sim 10$ in $Round_k$ is $19\%\sim 10\%$ higher than that in $Round_j$. For intuitively, the global model of $Round_k$ is more fair and effective than that of $Round_i$ and $Round_j$. However, the accuray variance of $Round_k$ is greater than that of $Round_i$ and $Round_j$. Nevertheless, by using our defination of fairness, the fairness of $Round_k$ is 0.116200599, while the fairness of $Round_i$ and $Round_j$ are 0.143478261 and 0.129881544 respectively. This shows that our definition of fairness can be well quantified the changes of fairness in the federated learning process, which will help us to understand the changes in the fairness of the global model in a timely manner.

3.3. The appropriate time to intervene

Most fairness federated learning methods give relatively large weights to the gradient updates of poor performance clients, so that the global model is biased to poor performance clients, so as to reduce the unfairness of the global model. For example, FedFa [5], q-FFL [6], and TERM [9]. In the $t+1$-th round, the weights applied to the gradient update of the client $k$of these three methods are as follows:

1. FedFa: $\alpha(-log_2 Acc_k^t/Acc)+\beta\left(-log_2\left(1-f_k^t/f\right)\right)$, where Acc_k^t is the training accuracy of the global model generated in round t on the data of client k, and f_k^t is the number of training participation of client k. We set $(\alpha,\beta)=(0.5,0.5)$.
2. q-FFL: $F_k^q(w^t)(q>1)$, where $F_k(w^t)$ is the training loss of the global model generated in round t on the data of client k;
3. TERM: $\frac{e^{tF_k(w^t)}}{\sum_{i=1}^{N}e^{tF_i(w^t)}}(t>0)$, where N is the number of selected clients in round t.

Although these works can alleviate the problem of performance unfairness in federated learning to a certain extent, they all ignore the dynamic of the process of federated learning. We still take Synthetic [21] as an example. Fig. 2 shows the degree of fairness intervention of different methods with the global iteration rounds of federated learning. The vertical axis is the sum of the distances between the weights of the clients, reflecting the degree of fairness intervention. As shown in Fig. 2, the degree of intervention of these methods in the early process of federated learning is significantly greater than that in the later stage. The reason is that in the early stage of federated learning, the model losses and accuracies of the initial global model on the clients are quite different; In the later stage of federated learning, the global model converges slowly, the difference between the model losses and accuracies of the global model on clients gradually decreases. However, in the early stage of federated learning, the change of the global gradient updates of two adjacent iterations is relatively large, and the updates by clients are similar, fairness intervention at this stage has little effect on improving fairness; On the contrary, in the later stage of federated learning, the gradient gap between the clients is larger, and the difference of the model performance of the clients becomes larger. However, the intervention degree of these methods in the process of federated learning in the early stage is greater than that in the later stage, which affects the true fairness of the model and even reduces the performance of the global model.

Definition 2. Global update scales G_s. Let $\Delta\mathbf{w}_{i,j}$define the (i,j)-th parameter in an update of the form $\Delta\mathbf{w}\in R^{p\times q}$, at some communication round t. For the sake of clarity, we will drop specific indexing of communication rounds for now. The parameter (i,j) in $\Delta\mathbf{w}$ is computed as $\Delta\mathbf{w}_{i,j}=\frac{1}{K}\sum_{k=1}^{K}\Delta\mathbf{w}_k^{i,j}$, where $\Delta\mathbf{w}_k^{i,j}$ is the (i,j)-th parameter in the update of $\Delta\mathbf{w}_k$. We then define the global update scales as the sum over all parameter variances in the updated matrix $\Delta\mathbf{w}$,

$$G_s=\frac{1}{p\times q}\sum_{i=0}^{p}\sum_{j=0}^{q}\Delta\mathbf{w}_{i,j}^2, \tag{7}$$

which represents the change of the global model in two iterations.

Table 1
Comparison of variance and Gini coefficient in federated learning.

	$Round_i$	$Round_j$	$Round_k$	improve
Client1	0.35	0.4165	0.495635	19%
Client2	0.4	0.472	0.55696	18%
Client3	0.45	0.5265	0.616005	17%
Client4	0.5	0.58	0.6728	16%
Client5	0.55	0.6325	0.727375	15%
Client6	0.6	0.684	0.77976	14%
Client7	0.65	0.7345	0.829985	13%
Client8	0.7	0.784	0.87808	12%
Client9	0.75	0.8325	0.924075	11%
Client10	0.8	0.88	0.968	10%
Variance	0.016666667	0.018035222	0.019146965	
Gini	0.143478261	0.129881544	0.116200599	

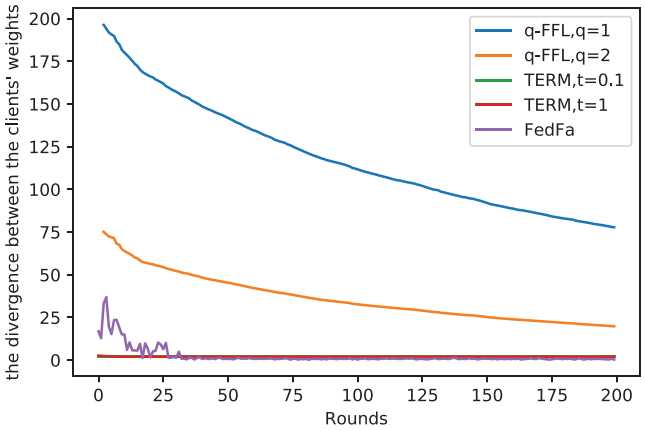


Fig. 2. The degree of fairness intervention of different methods with the global iteration rounds of federated learning.

Definition 3. Local dissimilarity. We define the variance of parameters of the client k as,

$$VAR[\Delta \mathbf{w}_k] = \frac{1}{p \times q} \sum_{i=0}^{p} \sum_{j=0}^{q} \left(\Delta \mathbf{w}_{i,j}^k - \Delta \mathbf{w}_{i,j} \right)^2. \tag{8}$$

We then define V_c as the sum parameter variances over all clients in the update matrix as,

$$V_c = \frac{1}{K} \sum_{k=0}^{K} VAR[\Delta \mathbf{w}_k]. \tag{9}$$

Further, the federated learning at $\mathbf{w}$ is said to be B-local dissimilarity:

$$B = \sqrt{\frac{V_c}{G_s}}. \tag{10}$$

B is a measure of dissimilarity between clients' updates.

As a sanity check, when all the local updates are the same, we have $B = 0$. However, in the federated setting, the data distributions are often heterogeneous and $B > 0$ due to sampling discrepancies even if the samples are assumed to be IID. The larger the value of B, the larger the dissimilarity among the local updates.

When should fairness intervention begin? Taking the results of FedAvg on Synthetic [21] as an example, we show the relationship between the global gradient update scales G_s and local dissimilarity B in Fig. 3. In order to show the results more intuitively, we select all clients to participate in the training in each round. As observable in Fig. 3, in the initial stage, the randomly initialized weights are greatly updated, so the gradient updates of two iterations of the global model change greatly. However, when approaching the local optimal value of the global model, the gradient updates of the global model will shrink sharply, the accuracy will converge, and the contributions of the clients will offset each other to a certain extent. Early intervention on the clients of federated learning will cause the updates of the global model to deviate from the real direction. We consider that if we start to intervene in fairness at the initial stage, the result is not ideal as the updates by clients are similar. In addition, we observe that the local dissimilarity rises with the number of iterations, as each client opti-

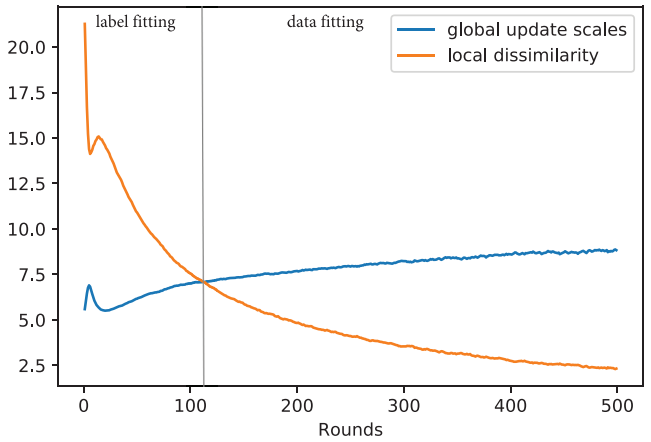


Fig. 3. The relationship between the global gradient update scales G_s and local dissimilarity B.

mizes on their local dataset. We take the closest point between the global update scales and the local dissimilarity as the separation point, with the label fitting phase on the left and the data fitting phase on the right. We propose that fairness intervention should be executed in the data fitting stage. At this stage, the global update scales change slowly and better reflect the real direction of the global model. Moreover, the local dissimilarity is also getting bigger, indicating that the unfairness is getting bigger. Intervene at this stage will not affect the real direction of the global model gradient update, but also achieve the desired fairness.

As only a small number of clients are selected to participate in the training in each iteration, the optimal dividing point cannot be obtained in advance. Therefore, we use the estimation method to obtain the approximate results. Firstly, we use the sliding window method to calculate the average value of the global update scale G_s in the window, and then calculate the difference of the global update scale G_s of two adjacent windows, which is recorded as ΔG_s. If the difference of ΔG_s corresponding to the two iterations is small, it indicates that the global update scale has become relatively slow, and fairness intervention can be carried out at this time. Assuming that the sliding window size is D, in t-th iteration, the server calculates the difference of ΔG_s between two adjacent windows as follows:

$$\Delta G_s^t = \frac{1}{D}\sum_{i=t-D}^{t} G_s^i - \frac{1}{D}\sum_{i=t-D-1}^{t-1} G_s^i < \eta. \tag{11}$$

The selection of η can be adjusted appropriately according to the sequence of ΔG_s in the previous period.

3.4. Fairness Federated Optimization Algorithm

In our work, we aim to obtain a global model $\mathbf{w}$ whose fairness meets our requirement while minimizing the average loss of all clients. The accuracy of the model depends on the specific form of the output function, such as SoftMax. On the contrary, it is more common to measure the unfairness of the losses of clients. To this end, we take the loss of the global model on each client k as an income for that client. Specifically, the bias of the global model toward any particular client is minimized by minimizing the inequality over the losses of the selected clients in a communication round. Therefore, the objective function is to minimize the following loss function:

$$F(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{w}) + \lambda G(F_1(\mathbf{w}), F_2(\mathbf{w}), \ldots, F_K(\mathbf{w})). \tag{12}$$

The first term is the average loss for all clients by the global model $\mathbf{w}$ after the update, while the second is the inequality of losses by the global model $\mathbf{w}$ before the update, and $F_1(\mathbf{w}) > F_2(\mathbf{w}) > \ldots > F_K(\mathbf{w})$. It is worth noting that the inequality measure is computed over a set of losses from the selected clients. Both terms are a function of the global model $\mathbf{w}$. We regard the loss of the global model for each client as an income for that client. Then for the federated network, its loss inequality over multiple clients is minimized to make the global model fairer. Specifically, in the federated learning process, the deviation of the global model to any specific client is minimized by minimizing the inequality of the losses over clients. Note that the first term of Eq. 12 is different from the objective function Eq. 2 in FedAvg, Eq. 12 does not consider the dataset size of clients. That is because FedAvg assumes that the target distribution is the uniform distribution of all clients' data. As this assumption is rather restrictive, it will lead to a biased global model toward the clients with a large amount of data. Therefore, in this paper, the objective function is to minimize the average loss of the global model on all clients' local training

data, regardless of the amount of data on the clients. And $\lambda$is a parameter that controls the proportion of the two terms. When $\lambda = 0$, that means no fairness is imposed on the personalized federated learning, which is much like [22]. A larger λ means that put more emphasis on fairness to reduce the difference in training accuracy between clients.

Then we solve the optimization problem through the gradient descent method. The gradients can be computed by:

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\Delta \mathbf{w}_k + \lambda \frac{\partial G(F_1(\mathbf{w}), \ldots, F_K(\mathbf{w}))}{\partial \mathbf{w}}. \tag{13}$$

$G(F_1(\mathbf{w}), \ldots, F_K(\mathbf{w}))$ is a composite function of $\mathbf{w}$, and its gradient is calculated as follows:

$$\begin{aligned} \tfrac{\partial G(F_1(\mathbf{w}),\ldots,F_K(\mathbf{w}))}{\partial \mathbf{w}} &= \tfrac{\partial G(F_1(\mathbf{w}))}{\partial F_1(\mathbf{w})}\tfrac{\partial F_1(\mathbf{w})}{\partial \mathbf{w}} + \ldots + \tfrac{\partial G(F_k(\mathbf{w}))}{\partial F_k(\mathbf{w})}\tfrac{\partial F_k(\mathbf{w})}{\partial \mathbf{w}} \\ &= \tfrac{1}{K}(\Delta \mathbf{w}_1 + 2\Delta \mathbf{w}_2 + \ldots + K(K-1)\Delta \mathbf{w}_K) \end{aligned} \tag{14}$$

Therefore, $\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}}$ is as follows:

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\Delta \mathbf{w}_k + \frac{1}{K}\sum_{k=1}^{K}\lambda k(k-1)\Delta \mathbf{w}_k. \tag{15}$$

The server updates the global model $\mathbf{w}$ as follows:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \alpha_t \frac{1}{N}\sum_{k=1}^{N}\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}}. \tag{16}$$

Different from the other methods, in which the clients' weights are designed empirically, we can obtain the precise relationship between the clients' gradient updates and the global model fairness through the gradient descent method.

From Eq. 15, we can see that the parameter λ control the amount of fairness. To avoid tuning this parameter, we design an adaptive parameter λ as follows:

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = \frac{\epsilon}{K}\sum_{k=1}^{K}\Delta \mathbf{w}_k + (1-\epsilon)\frac{k(k-1)}{\sum_{k=1}^{K}k(k-1)}\Delta \mathbf{w}_k. \tag{17}$$

As Eq. 17, the second term is the fairness penalty term. When we want greater fairness, that is, when ϵ is smaller, the second term is larger, which means that greater emphasis is placed on fairness; On the contrary, when ϵ becomes larger, the first term is larger, which means that the minimization of average loss is more emphasized.

Algorithm 1 illustrates the fairness federated optimization algorithm.

Algorithm 1: The fairness federated optimization algorithm

Input: $\mathbf{w}^{(0)}, \alpha$.
Output: $\mathbf{w}$.
for $t = 0$ to T **do**
  //Server:
  Select some clients ($k = \{1, \ldots, N\}$), and send the global model $\mathbf{w}^t$ to them;
  Waiting for and listening for signals from the selected client;
  Receive the local updates $\Delta \mathbf{w}_k$ and the loss $F_k(\mathbf{w})$ sent by clients;
  Calculate whether to start fairness intervention through Eq. (11);
  if start intervention **then**
    Update the global model $\mathbf{w}^{(t+1)}$ via Eq. (4);
  else
    Update the global model $\mathbf{w}^{(t+1)}$ via Eq. (16).
  end if
  //Client k, ($k = \{1, \ldots, N\}$):
  Receive the server's global model $\mathbf{w}$;
  Calculate the local updates $\Delta \mathbf{w}_k = \frac{\partial F_k(\mathbf{w})}{\partial \mathbf{w}}$;
  Send the current local updates $\Delta \mathbf{w}_k$ and the loss $F_k(\mathbf{w})$ to the server.
end for

4. Evaluation

In this section, we first introduce the experimental settings, including datasets, comparison methods, and experimental settings. Then we conduct experiments and analyses on the accuracy, fairness, and convergence of the model. We conduct ablation experiments on the proposed intervention time strategy, and finally, we analyze the experimental parameters.

4.1. Federated Datasets

(1) Synthetic [21]: For each client k, we generate a synthetise dataset (X_k, Y_k) based on model $y = argmax(softmax(Wx + b)$, where $x \in \mathbb{R}^{60}, W \in \mathbb{R}^{10\times 60}, b \in \mathbb{R}^{60}, W_k \sim N(\mu_k, 1), b_k \sim N(\mu_k, 1), \mu_k \sim N(0, \alpha), x_k \sim N(v_k, \sum)$, the diagonal of covariance matrix $\sum$ is $\sum_{i,j} = j^{-1.2}$. Each element in the average vector v_k comes from distribution $N(B_k, 1)$, where $B_k \sim N(0, \beta)$. Therefore, α controls the difference between the local models of clients, and β controls the difference between the local data on clients. "Synthetic-iid" means all devices follow the same data distribution, and "Synthetic-$\alpha\beta$" means heterogeneous distributed datasets generated by Synthetic(α, β). We generate a total of 30 clients, and the number of data samples on each client follows the power law. Our goal is to learn W and b.

(2) Sent140 [23]: Sent140 includes 1101 tweets accounts, of which each tweet account corresponds to a client. Our task is text emotion analysis, which we modeled as a binary classification problem and use the LSTM model for classification. The LSTM classifier includes two LSTM layers and one full connection layer. The model takes a sequence of 25 words as input, embeds each word into 300-dimensional space using Glove [24], and outputs a binary label.

(3) CIFAR-10 [25]: CIFAR-10 consists of 50, 000 training examples and 10, 000 testing examples from different 10 classes. These images are 32×32 pixels with three RGB channels. We generate 100 clients, and use the Dirichlet function to set different levels of Non-IID as prior work [14]. Cifar10-00 means that the local dataseta are I.I.D. drawn from the global distribution. Cifar10-05 means that the partitioned dataset obeys the dirichlet$(\alpha * 0.5)$ distirbution, and each client has a similar amount data size. In Cifar10-45, the data distribution is the same as Cifar10-05. However, the local data sizes vary across clients. We deploy experiments on AlexNet [26], which is a widely used CNNs, and contains 5 convolutional layers and 3 FC layers.

4.2. Compared methods

To verify the effectiveness and fairness of the proposed method, we compare it with FedAvg and other fairness methods:

1. q-FFL[6]: q-FFL regards the global model as a resource serving clients, and assigns greater weight to clients with greater loss, so as to reduce the model performance difference between clients. The weight it imposes on clients is the q-power of the local loss of these clients: $F_k^q(w^t)(q > 0)$, where $F_k(w^t)$ is the training loss of the global model in t-th round on client k. The greater the value of q, the greater the degree of fairness intervention.
2. FedFa[5]: Fedfa uses information theory to achieve fairness. It points out that the amount of information available to the client will vary according to the training accuracy and training frequency. For example, the lower the training accuracy, the more information the client needs to learn; The more times participate in training, the more information the client gets. Therefore, Fedfa gives higher weights to clients with lower training accuracy and more training times. FedFa(α, β) means the proportional effects of the frequency and training accuracy on the global model, in which α controls the weights of the training accuracy, and β controls the weights of the frequency.
3. TERM [9]: TERM adds a fairness penalty term to the average loss function of clients, and then uses the exponential smoothing method to trade-off between the average loss and maximum loss of clients. The weight it applies to the client is $\frac{e^{tF_k(w)}}{\sum_{i=1}^{N} e^{tF_i(w)}}(t > 0)$, where N is the number of clients selected, $F_k(w)$ is the training loss of the global model on client k. The greater the value of t, the greater the degree of fairness intervention.
4. FedProx [21]: To handle heterogeneous federated data, FedProx limits local model updates by penalizing large changes to the current model. FedProx is similar to FedAvg, but is more robust and has more stable convergence than FedAvg.
5. FedFV [14]: FedFV identify that conflicting gradients with large differences cause unfairness. It mitigates potential internal conflicts and external conflicts among clients before averaging their gradients. α represents the proportion of the clients that keep their original gradients, and τ controls the degree of mitigating external conflicts. We set $\alpha = 0.1, \tau = 10$.
6. FedAvg[3]: Fedavg is currently the most widely recognized federated averaging algorithm. The server collects gradient updates of clients and aggregates them into a new global model by averaging. The weight it imposes on the client is p_k, indicating the proportion of local data of the client k in all data $p_k = |D_k|/\sum_{k=1}^{K}|D_k|$. It shows that the weight applied by FedAvg has nothing to do with the model performance of clients, but only with the amount of data of clients.

4.3. Experimental Settings

For these datasets, we set the learning rate α as 0.01, 0.1 and 0.1 respectively, and randomly select 10 clients for training in each iteration. The batch sizes of the Synthetic, Sent140 and Cifar10 are 32, 64 and 64, respectively. We run 200, 400 and 600 iterations respectively on Synthetic, Sent140 and Cifar10, and select the results of the last 50 iterations for statistics.

4.4. Experiments Result Analysis

In this section, we conduct experiments and analyze results on the accuracy and fairness of the model. Table 1.

4.4.1. Accuracy

Fig. 4 shows the average testing accuracy of different methods with the number of iterations on Synthetic, Sent140 and Cifar10. Table 2 shows the comparison of testing accuracy and fairness of different methods, including the accuracy of the worst 10% client and the best 10% client.

From Fig. 4 and Table 2, we can see that the average testing accuracy of our proposed method FedGini is higher than other methods on Synthetic-00, Synthetic-05, Synthetic-11, Sent140, and Cifar10. And, on Synthetic-iid, FedGini's performance is also very close to that of FedFV. The average testing accuracy of the worst 10% clients has improved significantly: On Synthetic-11, Sent140, and Cifar10, FedGini is higher than other methods, and on Synthetic-00 and Synthetic-iid, the performance of the worst 10% clients are very close to that of the highest methods.

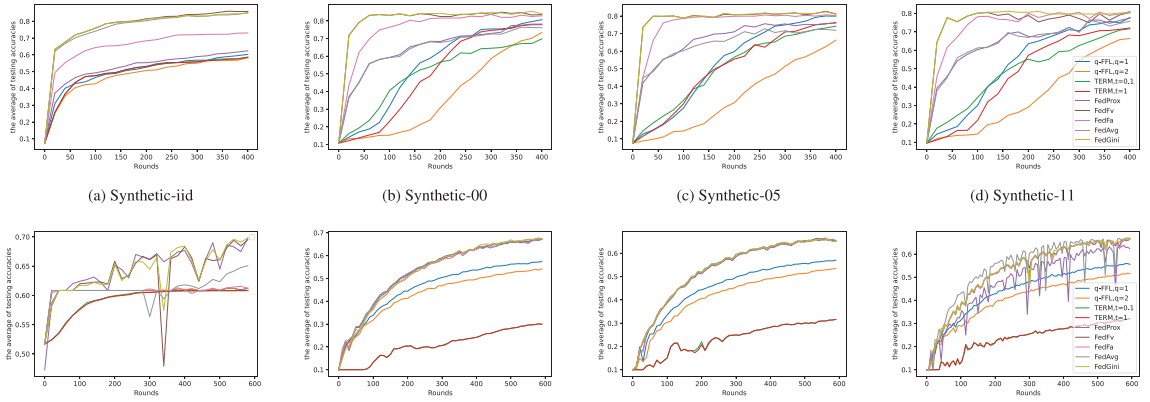
It is noted that the accuracy of q-FFL (q = 1) on Synthetic-00, Synthetic-05 and Synthetic-11 is a little higher that of FedAvg. And, the accuracy of q-FFL (q = 1) is lower than that of FedAvg on other datasets. This indicates that the fairness intervention imposed by q-FFL affects the convergence of the model. Especially when $q = 2$, the greater the degree of fairness intervention, the slower the convergence of the model. Similarly, the accuracy of TERM is higher than that of FedAvg on Synthetic-05, but lower on Synthetic-iid, Synthetic-11, Sent140, and Cifar10. As for Synthetic-00, TERM (T = 1) improves the accuracy than FedAvg, but TERM (T = 0.1) decreases the accuracy. This indicates that, the parameters have a great impact on the performance of the model. Sometimes, larger fairness parameters will affect the convergence and performance of the model.

We also found that on the IID datasets (Synthetic-iid and Cifar10-00), and the dataset with balanced data volume (Cifar10-05), FedAvg performes well. The fairness intervention on these datasets should be smaller, otherwise it may lead to the decline of model performance. For example, the performance of FedProx and FedFa on Synthetic-iid is lower than that of FedAvg. Our method FedGini and FedFV perform well on all datasets. The experimental results show that the fairness intervention degree imposed by our method FedGini is appropriate, which does not cause the decline of the model performance, but has a certain degree of improvement. This proves the robustness of FedGini.

4.4.2. Fairness

Fig. 5 shows the Gini coefficient of testing accuracy of different methods on Synthetic, Sent140 and Cifar10 with the number of iterations, which reflects the fairness of the federated learning process. Combined with Table 2, it can be seen that the Gini coefficient of FedGini's testing accuracy on Synthetic-00, Synthetic-05, Synthetic-11, Sent140, and Cifar10 is lower than other methods. And, on Synthetic-iid, FedGini's Gini coefficient is very close to that of FedFV. This indicates that FedGini will lead to more fair model accuracy.

We notices that the Gini coefficient curves of TERM (t = 0.1) and TERM (t = 1) has a cross-point on Synthetic. Similarly, the Gini coefficient curves of q-FFL (q = 1) and q-FFL (q = 2) has a cross-point on Synthetic-00 and Synthetic-11. This shows



(a) Synthetic-iid (b) Synthetic-00 (c) Synthetic-05 (d) Synthetic-11

(e) Sent140 (f) Cifar10-00 (g) Cifar10-05 (h) Cifar10-45

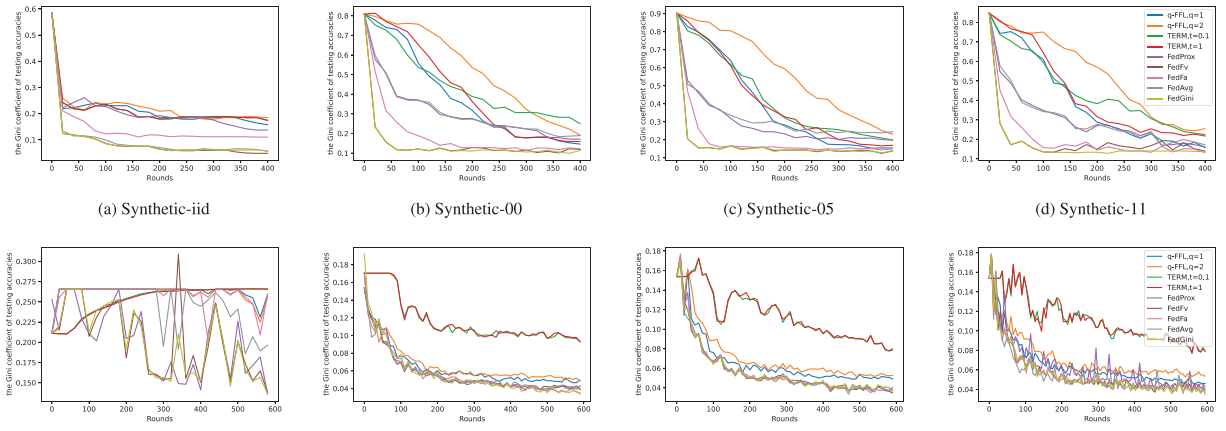
Fig. 4. The average testing accuracy.

Table 2
Comparison of model accuracy and fairness of different methods.

Datasets	Methods		Accuracy	Worst 10%	Best 10%	Variance	Gini
Synthetic-iid	q-FFL	q = 1	59.39%±0.34	28.89%±2.55	97.78%±0.00	327±17	0.167±0.005
		q = 2	56.89%±0.36	19.44%±0	96.27%±1.04	401±6	0.188±0.001
	TERM	t = 0.01	57.82%±0.44	21.44%±2.67	97.78%±0.00	385±24	0.181±0.005
		t = 1	57.70%±0.40	20.67%±2.30	97.78%±0.00	392±22	0.182±0.004
	FedProx		61.54%±0.47	34.72%±0	97.78%±0.00	260±4	0.140±0.003
	FedFV		**85.74%±0.16**	**71.25%±1.31**	**100%±0.00**	**61±6**	**0.049±0.002**
	FedFa	$\alpha = 0, \beta = 1$	73.27%±0.27	43.45%±0	94.07%±0.00	207±4	0.108±0.002
		$\alpha = 0.5, \beta = 0.5$	72.65%±0.35	43.45%±0	94.07%±0.00	210±1	0.111±0.001
		$\alpha = 1, \beta = 0$	72.26%±0.10	43.45%±0	94.07%±0.00	209±4	0.111±0.000
	FedAvg		84.26%±0.54	67.36%±1.48	100%±0.00	81±9	0.062±0.004
	FedGini		84.48%±0.36	67.26%±1.04	100%±0.00	84±6	0.059±0.002
Synthetic-00	q-FFL	q = 1	78.96%±1.13	17.22%±1.36	100%±0	613±40	0.158±0.008
		q = 2	70.58%±1.69	7.52%±5.97	100%±0	855±118	0.220±0.018
	TERM	t = 0.01	67.92%±1.12	0%±0	100%±0	1255±79	0.274±0.014
		t = 1	77.28%±0.49	13.89%±0	100%±0	622±11	0.163±0.004
	FedProx		77.82%±0.42	9.56%±2.49	100%±0	755±40	0.174±0.005
	FedFV		84.24%±0.55	**43.5%±4.93**	100%±0	341±35	0.109±0.006
	FedFa	$\alpha = 0, \beta = 1$	82.62%±0.25	34.89%±3.25	100%±0	393±32	0.121±0.004
		$\alpha = 0.5, \beta = 0.5$	82.93%±0.32	35.17%±3.17	100%±0	394±29	0.121±0.004
		$\alpha = 1, \beta = 0$	83.08%±0.24	35.94%±2.51	100%±0	386±21	0.119±0.003
	FedAvg		76.39%±0.88	10.33%±1.93	100%±0	785±47	0.183±0.008
	FedGini		**84.61%±0.39**	42.44%±2.89	100%±0	**336±19**	**0.106±0.004**
Synthetic-05	q-FFL	q = 1	79.49%±0.38	22.22%±0	100%±0	590±12	0.156±0.002
		q = 2	61.50%±1.77	14.21%±1.37	100%±0	815±44	0.263±0.015
	TERM	t = 0.01	73.01%±0.55	11.11%±0	100%±0	850±21	0.212±0.005
		t = 1	75.34%±0.42	22.22%±0	100%±0	573±11	0.169±0.003
	FedProx		75.56%±0.34	13.41%±2.55	100%±0	863±37	0.203±0.005
	FedFV		81.20%±0.83	37.75%±2.84	100%±0	453±49	0.140±0.009
	FedFa	$\alpha = 0, \beta = 1$	79.66%,0.40	31.59%±3.11	100%±0	555±38	0.157±0.005
		$\alpha = 0.5, \beta = 0.5$	80.30%±0.34	34.92%±1.50	100%±0	501±22	0.149±0.004
		$\alpha = 1, \beta = 0$	80.12%±0.38	33.21%±1.80	100%±0	518±28	0.151±0.004
	FedAvg		71.90%±0.67	1.14%±0	100%±0	798±29	0.242±0.006
	FedGini		**81.20%±0.84**	**37.78%±2.86**	100%±0	**453±49**	**0.140±0.009**
Synthetic-11	q-FFL	q = 1	75.96%±1.14	33.32%±2.99	100%±0	537±44	0.169±±0.009
		q = 2	64.77%±1.43	7.27%±3.55	100%±0	861±40	0.252±0.011
	TERM	t = 0.01	69.84%±1.48	5.68%±0.61	100%±0	915±25	0.236±0.010
		t = 1	71.34%±0.90	14.70%±3.59	100%±0	826±43	0.220±0.008
	FedProx		76.49%±0.67	29.89%±4.14	100%±0	565±31	0.170±0.005
	FedFV		78.07%±1.92	32.80%±7.48	100%±0	546±105	0.163±0.019
	FedFa	$\alpha = 0, \beta = 1$	79.91%±0.61	42.44%±3.81	100%±0	451±39	0.144±0.006
		$\alpha = 0.5, \beta = 0.5$	79.91%±0.71	42.11%±4.39	100%±0	452±45	0.144±0.007
		$\alpha = 1, \beta = 0$	79.97%±0.77	42.61%±5.17	100%±0	448±50	0.143±0.008
	FedAvg		74.26%±1.08	22.47%±1.62	100%±0	613±35	0.188±0.009
	FedGini		**80.08%±0.25**	**44.44%±0**	100%±0	**410±3**	**0.139±0.001**
Sent140	q-FFL	q = 1	61.20%±0.11	11.08%±1.31	99.84%±0.56	734±51	0.252±0.009
		q = 2	61.15%±0.18	11.79%±1.50	99.67%±0.83	706±60	0.247±0.011
	TERM	t = 0.01	60.84%±0.09	8.71%±0.00	100%±0.00	801±2	0.266±0.001
		t = 1	60.82%±0.09	8.71%±0.00	100%±0.00	802±1	0.266±0.001
	FedProx		66.79%±3.17	28.52%±8.35	98.34%±1.90	432±137	0.175±0.034
	FedFV		67.06%±2.89	29.08%±8.40	98.38%±1.84	426±140	0.173±0.034
	FedFa	$\alpha = 0, \beta = 1$	61.15%±0.16	10.26%±1.36	99.95%±0.19	761±44	0.257±0.008
		$\alpha = 0.5, \beta = 0.5$	61.43%±0.24	12.70%±2.64	99.48%±1.20	690±87	0.243±0.017
		$\alpha = 1, \beta = 0$	61.48%±0.46	14.20%±3.11	98.99%±1.72	647±100	0.234±0.019
	FedAvg		63.85%±1.33	22.04%±6.03	97.88%±2.62	805±117	0.198±0.029
	FedGini		**67.24%±2.88**	**29.89%±7.89**	98.23%±1.90	**411±131**	**0.169±0.032**
Cifar10-00	q-FFL	q = 1	57.18%±0.32	48.49%±0.44	66.41%±0.43	25±1	0.049±0.0015
		q = 2	53.87%±0.30	45.02%±0.32	62.46%±0.34	24±0.8	0.051±0.0011
	TERM	t = 0.01	29.88%±0.15	21.88%±0.26	38.91%±0.28	26±1	0.096±0.0022
		t = 1	29.88%±0.17	21.84%±0.29	38.89%±0.26	26±1	0.097±0.0022
	FedProx		66.46%±0.38	57.87%±0.60	74.61%±0.68	23±2	0.041±0.0020
	FedFV		66.80%±0.39	58.62%±0.67	74.94%±0.53	22±2	0.040±0.0019
	FedFa	$\alpha = 0, \beta = 1$	67.37%±0.43	59.70%±0.73	74.77%±0.73	19±2	0.036±0.002
		$\alpha = 0.5, \beta = 0.5$	67.19%±0.45	59.86%±0.91	74.84%±0.58	19±2	0.036±0.0020
		$\alpha = 1, \beta = 0$	67.23%±0.46	59.48%±0.76	74.25%±0.57	18±2	0.036±0.002
	FedAvg		66.75%±0.37	58.74%±0.57	74.91%±0.71	23±2	0.040±0.0021
	FedGini		**67.43%±0.32**	**59.58%±0.64**	**74.55%±0.45**	**18±2**	**0.035±0.002**
Cifar10-05	q-FFL	q = 1	56.85%±0.24	48.58%±0.37	65.93%±0.36	26±1	0.050±0.001
		q = 2	53.18%±0.37	44.51%±0.53	62%±0.32	25±1	0.053±0.001

Table 2 (*continued*)

Datasets	Methods		Accuracy	Worst 10%	Best 10%	Variance	Gini
	TERM	t = 0.01	31.30%±0.31	23.55%±0.43	39.24%±0.29	20±1	0.081±0.003
		t = 1	31.30%±0.32	23.48%±0.45	39.21%±0.29	20±1	0.082±0.003
	FedProx		65.76%±0.46	57.79%±0.76	73.38%±0.69	20±2	0.039±0.002
	FedFV		65.93%±0.40	58.11%±0.67	73.48%±0.75	20±2	0.038±0.002
	FedFa	$\alpha = 0, \beta = 1$	65.73%±0.43	56.89%±0.79	73.38%±0.75	23±2	0.041±0.002
		$\alpha = 0.5, \beta = 0.5$	65.97%±0.44	57.87%±0.67	73.20%±0.66	19±2	0.038±0.002
		$\alpha = 1, \beta = 0$	66.04%±0.44	58.04%±0.79	73.60%±0.70	21±2	0.038±0.002
	FedAvg		65.98%±0.47	58.01%±0.76	73.71%±0.68	20±2	0.039±0.002
	FedGini		**66.50%±0.39**	**58.65%±0.55**	**74.06%±0.71**	**21±1**	**0.038±0.002**
Cifar10-45	q-FFL	q = 1	55.56%±0.34	47.36%±0.49	63.61%±0.52	22±1	0.0477±0.001
		q = 2	51.26%±0.39	42.39%±0.39	59.70%±0.47	24±1	0.054±0.001
	TERM	t = 0.01	31.57%±0.44	23.74%±0.42	39.35%±0.36	**20±1**	0.080±0.003
		t = 1	31.58%±0.42	23.73%±0.46	39.36%±0.38	**20±1**	0.081±0.003
	FedProx		62.30%±3.73	53.69%±3.68	70.60%±3.72	24±3	0.044±0.004
	FedFV		65.79%±0.92	57.10%±1.23	73.96%±0.88	23±2	0.041±0.002
	FedFa	$\alpha = 0, \beta = 1$	65.73%±0.77	57.64%±1.12	73.92%±0.99	22±2	0.040±0.002
		$\alpha = 0.5, \beta = 0.5$	65.93%±0.84	57.55%±1.16	**73.99%±0.86**	22±2	0.040±0.002
		$\alpha = 1, \beta = 0$	65.85%±1.02	57.21%±1.04	74.01%±1.14	23±3	0.041±0.002
	FedAvg		65.55%±2.98	56.63%±3.03	73.18%±2.96	22±2	0.041±0.003
	FedGini		**66.17%±0.60**	**58.10%±0.88**	73.97%±0.59	21±2	**0.039±0.002**



(a) Synthetic-iid (b) Synthetic-00 (c) Synthetic-05 (d) Synthetic-11

(e) Sent140 (f) Cifar10-00 (g) Cifar10-05 (h) Cifar10-45

Fig. 5. The Gini coefficient of testing accuracy.

that in the stage before the cross-point, the methods with larger fairness intervention converge more slowly, while in the stage after the cross-point, these methods will have a faster convergence speed. In addition, the performance of TERM (t = 0.1) and TERM (t = 1) on the Sent140 is very similar: the Gini coefficient changes greatly in the first 10 rounds and basically doesn't change after the 10th round, indicating that the fair intervention in the early stage was too large and the intervention in the later stage was too small. These experimental results show that the fairness intervention in the early stage will slow down the convergence speed of the model, and the greater the degree of intervention, the slower the convergence; Fairness intervention in the later stage of federated learning can improve the convergence speed, and the greater the degree of intervention, the faster the convergence speed. Given all of that, we can see that the fairnesses of q-FFL and TERM in different datasets are quite different, and the influence of parameters is also great, indicating that these methods are not suitable for different datasets, and the selection of parameters is difficult. Our method FedGini performs best on Synthetic, Sent140 and Cifar10, which shows that FedGini is more stable.

By comparing the Gini coefficient and variance, we find that the variance and Gini coefficient of q-FFL (q = 2) on Synthetic-11 are 815 and 0.263 respectively, while that of TERM (t = 0.01) are 850 and 0.212 respectively. The testing accuracy of q-FFL (q = 2) and TERM (t = 0.01) on Synthetic-11 are 61.50 and 73.01 respectively. We can see that the worse the model performance is, the smaller the variance is, so the variance is not suitable to compare the models with different per-

formances. Conversely, the Gini coefficient can quantify the fairness of the federated learning process, and can be applied to compare the fairness of models with different performances.

4.5. Ablation Experiment

To illustrate the role of intervention time strategy, we set up two comparison methods: One is that FedGini conducts fairness intervention from the beginning; The other is that FedGini only intervenes in the data fitting stage. We set $\epsilon = 0$, and the data fitting stage on Synthetic-11 is after the 100th round, the data fitting stage on Sent140 is after the 220th round, and the data fitting stage on Cifar10-05 is after the 400th round. Fig. 6 and Fig. 7 respectively show the average training accuracy and Gini coefficient of training accuracy of the two methods on different datasets.

It can be seen from Fig. 6a and Fig. 7a that the convergence speed of FedGini with intervention time strategy on Synthetic-11 is much faster than that of FedGini with fairness intervention from the beginning, and the curve is smoother, indicating that the model performance and fairness are more stable. This proves that the fairness intervention in the early stage of federated learning will slow down the convergence speed.

From Fig. 6b, Fig. 7b, Fig. 6c, and Fig. 7c, we can also see that FedGini with intervention time strategy has faster convergence speed on Sent140 and Cifar10-05. In addition, we found that in Fig. 6b, FedGini, which conducts fairness intervention from the beginning, only changes the accuracy in the first 20 rounds, and the accuracy curve after 20 rounds is unchanged, indicating that the fairness intervention in the early stage of federated learning is too large, which seriously damages the performance of the model. In Fig. 7b, the fairness of FedGini who intervened in fairness from the beginning is no better than that of FedGini who intervened later, which shows that the fairness intervention in the early stage of federated learning can not achieve the desired fairness.

In conclusion, it shows that the intervention time strategy can accelerate the convergence speed of the model and lead to better fairness.

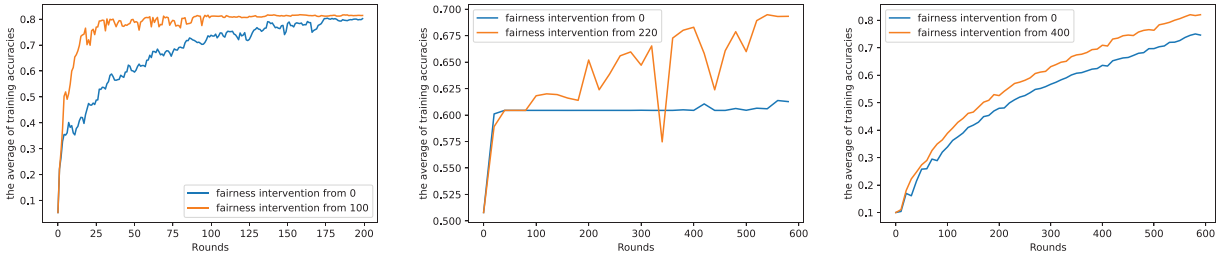
4.6. Parameter Analysis

FedGini only needs to adjust ϵ, which is the expected fairness. $\epsilon$is a Gini coefficient, so it ranges from 0 to 1. The smaller the ϵ, the greater the expected fairness. Accordingly, the larger ϵ, the smaller the expected fairness. When $\epsilon = 1.0$, it means that FedGini does not consider fairness. Note that FedGini ($\epsilon = 1.0$) is different from FedAvg, FedGini does not consider the weight of the clients' dataset size for fairness. Table 3 summarizes the testing accuracy and model fairness under different fairness constraints. Fig. 8 and Fig. 9 respectively show the average training accuracy and the Gini coefficient of training accuracy of FedGini under different fairness constraints.

First of all, we found that the testing accuracy difference and the Gini coefficient of testing accuracy under different parameters are small, that is, the FedGini can achieve good performance under any parameters, which proves the robustness of the proposed method.

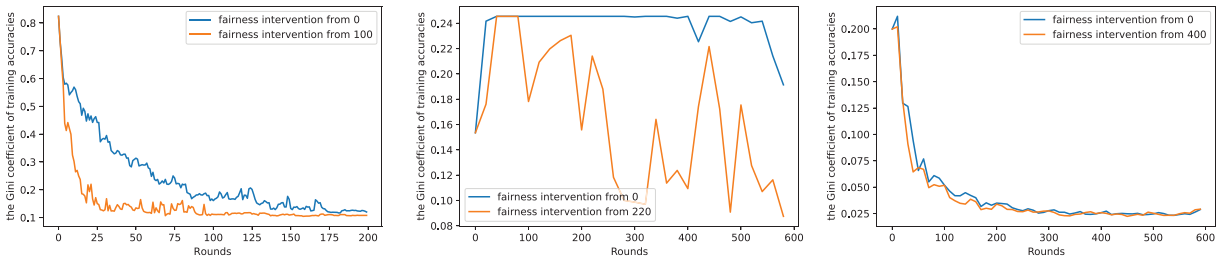
Then, from Table 3, we can see that FedGini ($\epsilon = 0.0$) has smaller Gini coefficient of testing accuracy than FedGini ($\epsilon = 1.0$), and the worst 10% clients have higher testing accuracy on both Synthetic-11 and Sent140. Moreover, the average testing accuracy of FedGini ($\epsilon = 0.0$) is lower than that of FedGini ($\epsilon = 0.5$) on Sent140 and Cifar10-05. This shows that FedGini ($\epsilon = 0.0$) can improve the fairness of the model. However, the performance of the model may be degraded.

From Fig. 8 and Fig. 9, we can see that, when $\epsilon = 1.0$, the fairness and training accuracy of FedGini fluctuate most violently, while when $\epsilon = 0.0$, the fairness and training accuracy are the most stable. The jitter of the curve reflects the difference in training accuracy and fairness of the global model of two adjacent iterations. In federated learning, the number of clients participating in each round is less than the total number of clients. The curve jitter is strong, indicating that the model is more inclined to the clients participating in the training, as the clients participating in each round are different, the training accuracy and fairness change too much. Fairness intervention can reduce the dependence of the model on the participating clients, because if the client performs well, it will be given a lower weight. It can be seen that the smaller the ϵ, that is, the greater the fairness constraint, the more stable the accuracy and fairness of the model.



(a) Synthetic-11 (b) Sent140 (c) Cifar10-05

Fig. 6. The average training accuracy under different intervention time.

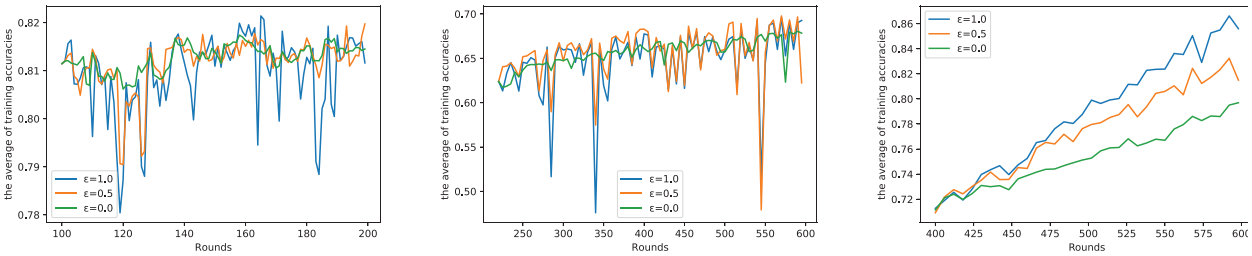


(a) Synthetic-11 (b) Sent140 (c) Cifar10-05

Fig. 7. The Gini coefficient of training accuracy under different intervention time.

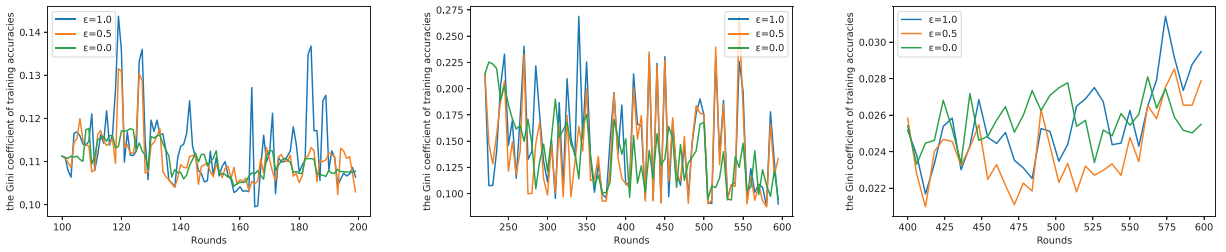
Table 3
Training accuracy and fairness results under different fairness constraints.

Datesets	ϵ	Accuracy	Worst 10%	Best 10%	Variance	Gini
Synthetic-11	$\epsilon = 1.0$	79.01%±0.47	36.63%±3.89	100%±0	502±38	0.154±0.006
	$\epsilon = 0.5$	79.02%±0.50	37.38%±4.41	100%±0	501±39	0.153±0.006
	$\epsilon = 0.0$	**79.17%±0.61**	**39.31%±5.78**	**100%±0**	**484±52**	**0.151±0.008**
Sent140	$\epsilon = 1.0$	66.91%±3.67	29.52%±7.60	**98.26%±1.85**	412±115	0.171±0.032
	$\epsilon = 0.5$	**67.24%±2.88**	29.89%±7.89	98.23%±1.90	**411±131**	0.169±0.032
	$\epsilon = 0.0$	67.18%±1.24	**31.36%±4.70**	97.82%±1.62	424±78	**0.163±0.018**
Cifar10-05	$\epsilon = 1.0$	66.23%±0.53	58.22%±0.74	73.89%±0.67	21±2	0.039±0.002
	$\epsilon = 0.5$	**66.50%±0.39**	**58.65%±0.55**	**74.06%±0.71**	21±1	0.038±0.002
	$\epsilon = 0.0$	65.41%±0.38	57.88%±0.56	72.72%±0.67	**20±2**	**0.038±0.002**



(a) Synthetic-11 (b) Sent140 (c) Cifar10-05

Fig. 8. Average training accuracy under different fairness constraints.



(a) Synthetic-11 (b) Sent140 (c) Cifar10-05

Fig. 9. The Gini coefficient of training accuracy under different fairness constraints.

5. Conclusion

In this paper, we study the fairness problem for heterogeneous federated learning. Firstly, we use Gini coefficient to quantify the fairness of federated learning, which can reflect the fairness of federated learning process in each round; Secondly, different from the previous methods, we divide federated learning into two different stages: label fitting and data fitting, and propose fairness intervention in the data fitting stage, which will not affect the real direction of global model update, but also

achieve the desired fairness; Then, we propose a fairness federated optimization algorithm, which introduces the fairness penalty term into the objective function, and obtains the relationship between the clients' gradient updates and the global model fairness through gradient descent, so as to improve the fairness of the model while maintaining the performance of the global model; Finally, in order to evaluate the effectiveness and fairness of our method, we conduct experiments on Synthetic, Sent140 and Cifar10 respectively, and the experimental results prove the effectiveness and fairness of our method.

However, our method also has some shortcomings. For example, for the dataset with fast convergence, our intervention time strategy has little effect because there is no obvious boundary between label fitting and the data fitting stage. In addition, we cannot find the optimal dividing point of label fitting and data fitting stage in the case that only a small number of clients are selected to participate in the training in each iteration. We will continue to discuss this issue in-depth in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research is supported by the National Key R&D Program of China (No. 2020YFB1006001), the National Natural Science Foundation of China (62176269, 62076179, 61732011), the Beijing Natural Science Foundation (Z180006), the Innovative Research Foundation of Ship General Performance (25622112), the Natural Science Foundation of Hubei Province in China (2021CFB482), and Training fund for teachers' scientific research ability (2022pygpzk05).

Appendix A. Convergence Analysis

The objective function of the global model is as follows:

$$F(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{w}) + \lambda G(F_1(\mathbf{w}), F_2(\mathbf{w}), \ldots, F_K(\mathbf{w})). \tag{A.1}$$

The gradient of the global model is calculated as follows:

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = \frac{\epsilon}{K}\sum_{k=1}^{K}\Delta\mathbf{w}_k + (1-\epsilon)\frac{k(k-1)}{\sum_{k=1}^{K}k(k-1)}\Delta\mathbf{w}_k. \tag{A.2}$$

We use M_k to represent $\frac{k(k-1)}{\sum_{k=1}^{K}k(k-1)}$, then Eq. (A.2) can be rewritten as follows:

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = \frac{\epsilon}{K}\sum_{k=1}^{K}\Delta\mathbf{w}_k + (1-\epsilon)M_k\Delta\mathbf{w}_k. \tag{A.3}$$

Assumption 1. The objective function of all clients $F_k, (k = \{1, 2, 3, \ldots, K\})$ are L-smooth, that is, for all model parameter matrices $\mathbf{v}$ and $\mathbf{w}$, the following inequality holds:

$$F_k(\mathbf{v}) \leqslant F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T\Delta F_k(\mathbf{w}) + \frac{L}{2}\|\mathbf{v}-\mathbf{w}\|_2^2. \tag{A.4}$$

Assumption 2. The objective function of all clients $F_k, (k = \{1, 2, 3, \ldots, K\})$ are μ-strongly convex, that is, for all model parameter matrices $\mathbf{v}$ and $\mathbf{w}$, the following inequality holds:

$$F_k(\mathbf{v}) \geqslant F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T\Delta F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|_2^2. \tag{A.5}$$

Assumption 3. For each client k, $(k \in \{1, 2, 3, \ldots, K\})$, in each iteration t, $(t \in \{1, 2, 3, \ldots, T\})$, the expected squared norm of stochastic local gradients is bounded:

$$\mathbb{E}\left\|\Delta\mathbf{w}_k^t\right\|^2 \leqslant G^2. \tag{A.6}$$

Definition 1. We use the following formula to quantify the degree of data heterogeneity in federated learning:

$$\Gamma = F^* - \sum_{k=1}^{K} F_k^*. \tag{A.7}$$

where F^* is the optimal function value of the global objective function, F_k^* is the optimal function value obtained by optimizing the client k. If the data is IID, then γwill become zero as the number of data samples increases. If the data is Non-IID, that is, heterogeneous, $\gamma$is non-zero, and its size reflects the heterogeneity of data.

Lemma 1. In t-th iteration, the distance between the local parameter matrix $\mathbf{w}_k^t$ and the global parameter matrix $\mathbf{w}^t$ is bounded:

$$\mathbb{E}\sum_{k=1}^{K}\left\|\mathbf{w}_t - \mathbf{w}_k^t\right\|^2 \leqslant (t-t_0)^2\alpha^2 G^2. \tag{A.8}$$

Proof 1.

$$\begin{aligned}\mathbb{E}\sum_{k=1}^{K}\left\|\mathbf{w}_t - \mathbf{w}_k^t\right\|^2 &= \mathbb{E}\sum_{k=1}^{K}\left\|\left(\mathbf{w}_k^t - w_{t_0}\right) - \left(w_t - w_{t_0}\right)\right\|^2 \\ &\leqslant \mathbb{E}\sum_{k=1}^{K}\left\|\mathbf{w}_k^t - w_{t_0}\right\|^2\end{aligned} \tag{A.9}$$

By Jensen inequality, we have:

$$\begin{aligned}\left\|\mathbf{w}_k^t - w_{t_0}\right\|^2 &= \left\|\sum_{t=t_0}^{t-1}\alpha\Delta\mathbf{w}_k^t\right\|^2 \\ &\leqslant (t-t_0)\sum_{t=t_0}^{t-1}\alpha^2\left\|\Delta\mathbf{w}_k^t\right\|^2\end{aligned} \tag{A.10}$$

Plugging Eq. (A.10) into Eq. (A.9), we have:

$$\begin{aligned}\mathbb{E}\sum_{k=1}^{K}\left\|\mathbf{w}_t - \mathbf{w}_k^t\right\|^2 &\leqslant \sum_{k=1}^{K}\mathbb{E}\sum_{t=t_0}^{t-1}(t-t_0)\alpha^2\left\|\Delta\mathbf{w}_k^t\right\|^2 \\ &\leqslant \sum_{k=1}^{K}\sum_{t=t_0}^{t-1}(t-t_0)\alpha^2 G^2 \\ &\leqslant \sum_{k=1}^{K}(t-t_0)^2\alpha^2 G^2 \\ &\leqslant (t-t_0)^2\alpha^2 G^2\end{aligned} \tag{A.11}$$

Lemma 2. Results of one step gradient descent:

$$\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2 \leqslant (1-\mu\alpha)\left\|\mathbf{w}^t - \mathbf{w}^*\right\|^2 + 2\sum_{k\in S^t}\left\|\mathbf{w}^t - \mathbf{w}_k^t\right\|^2 + 6\alpha^2 L\Gamma. \tag{A.12}$$

Proof 2.

$$\begin{aligned}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|^2 &= \left\|\mathbf{w}^t - \alpha\sum_{k\in S^t}\Delta F_k\left(\mathbf{w}_k^t\right) - \mathbf{w}^*\right\|^2 \\ &= \left\|\mathbf{w}^t - \mathbf{w}^*\right\|^2 - 2\alpha\sum_{k\in S^t}\Delta F_k\left(\mathbf{w}_k^t\right), \mathbf{w}^t - \mathbf{w}^* + \alpha^2\left\|\sum_{k\in S^t}\Delta F_k\left(\mathbf{w}_k^t\right)\right\|^2\end{aligned} \tag{A.13}$$

Plugging Eq. (A.3) into Eq. (A.13), we have:

$$\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2 \leqslant \left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 - 2\alpha\epsilon\sum_{k\in S^t}\Delta\mathbf{w}_k^t, \mathbf{w}^t-\mathbf{w}^* - 2\alpha(1-\epsilon)M_k\Delta\mathbf{w}_k^t, \mathbf{w}^t-\mathbf{w}^* + \alpha^2\epsilon^2\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 + \alpha^2(1-\epsilon)^2\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 \tag{A.14}$$

By Cauchy-Schwartz inequality and AM-GM inequality, we have:

$$-2\Delta F_k(\mathbf{w}_k^t), \mathbf{w}^t-\mathbf{w}_k^t \leqslant \frac{1}{\alpha}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 + \alpha\left\|\Delta F_k(\mathbf{w}_k^t)\right\|^2. \tag{A.15}$$

As $F_k(\cdot)$ is μ-strongly convex, we have:

$$-\Delta F_k(\mathbf{w}_k^t), \mathbf{w}_k^t-\mathbf{w}^* \leqslant -\frac{1}{\mu}\left\|\mathbf{w}_k^t-\mathbf{w}^*\right\|^2 - \left(F_k(\mathbf{w}_k^t)-F_k(\mathbf{w}^*)\right). \tag{A.16}$$

As $F_k(\cdot)$ is L-smooth, we have:

$$\left\|\Delta F_k(\mathbf{w}_k^t)\right\|^2 \leqslant 2L\left(F_k(\mathbf{w}_k^t)-F_k^*\right). \tag{A.17}$$

By Cauchy-Schwartz inequality and AM-GM inequality, we have:

$$\begin{aligned} -2M_k\Delta\mathbf{w}_k^t, \mathbf{w}^t-\mathbf{w}^* &\leqslant \alpha\left\|M_k\Delta\mathbf{w}_k^t\right\|^2 + \tfrac{1}{\alpha}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 \\ &\leqslant \alpha\left\|\Delta\mathbf{w}_k^t\right\|^2 + \tfrac{1}{\alpha}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2. \end{aligned} \tag{A.18}$$

Plugging Eq. (A.15), Eq. (A.16), Eq. (A.17), and Eq. (A.18) into Eq. (A.14), we have:

$$\begin{aligned} \left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2 &\leqslant \left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 + \varepsilon\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 + \alpha^2\varepsilon\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 \\ &\quad +(1-\varepsilon)\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 + \alpha^2(1-\varepsilon)\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 \\ &\quad +\alpha^2\epsilon^2\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 + \alpha^2(1-\epsilon)^2\left\|\sum_{k\in S^t}\Delta\mathbf{w}_k^t\right\|^2 \\ &= \left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 + 2\alpha^2\left\|\Delta F_k(\mathbf{w}_k^t)\right\|^2 + \left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 \\ &\leqslant \left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 + 2L\alpha^2\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F_k^*\right) \\ &\quad +\alpha\sum_{k\in S^t}\left(\tfrac{1}{\alpha}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 + \alpha\left\|\Delta F_k(\mathbf{w}_k^t)\right\|^2\right) \\ &\quad -2\alpha\sum_{k\in S^t}\left(\tfrac{1}{\mu}\left\|\mathbf{w}_k^t-\mathbf{w}^*\right\|^2 + \left(F_k(\mathbf{w}_k^t)-F_k(\mathbf{w}^*)\right)\right. \quad = (1-\mu\alpha)\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2 + \sum_{k\in S^t}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2 \\ &\quad \left. +\underbrace{4L\alpha^2\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F_k^*\right) - 2\alpha\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F_k(\mathbf{w}^*)\right)}_{A}\right). \end{aligned} \tag{A.19}$$

We next aim to bound A. We define $\gamma = 2\alpha(1-2L\alpha)$, where $\alpha \leqslant \frac{1}{4L}$, and $\alpha \leqslant \gamma_t \leqslant 2\alpha$. We split A_1 into two terms:

$$\begin{aligned} A &= -2\alpha(1-2L\alpha)\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F_k^*\right) + 2\alpha\sum_{k\in S^t}\left(F_k\left(\overset{*}{\mathbf{w}}\right)-F_k^*\right) \\ &= -\gamma\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F^*\right) + (2\alpha-\gamma)\sum_{k\in S^t}(F^*-F_k^*) \\ &= \underbrace{-\gamma\sum_{k\in S^t}\left(F_k\left(\overset{t}{\mathbf{w}}_k\right)-F^*\right)}_{B} + 4L\alpha^2\Gamma \end{aligned} \tag{A.20}$$

As $F_k(\cdot)$ is μ-strongly convex and L-smooth, by AM-GM inequality, we have:

$$\begin{aligned}\sum_{k\in S^t}\left(F_k(\mathbf{w}_k^t)-F^*\right) &= \sum_{k\in S^t}\left(F_k(\mathbf{w}_k^t)-F_k(\mathbf{w}^t)\right)+\sum_{k\in S^t}\left(F_k(\mathbf{w}^t)-F^*\right)\\ &\geqslant \sum_{k\in S^t}\left\langle\Delta F_k(\mathbf{w}^t),\mathbf{w}_k^t-\mathbf{w}^t\right\rangle+\left(F(\mathbf{w}^t)-F^*\right)\\ &\geqslant -\frac{1}{2}\sum_{k\in S^t}\left[\frac{1}{\alpha}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2+\alpha\left\|\Delta F_k(\mathbf{w}_k^t)\right\|^2\right]+\left(F(\mathbf{w}^t)-F^*\right)\\ &\geqslant -\sum_{k\in S^t}\left[\frac{1}{2\alpha}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2+\alpha L\left(F_k(\mathbf{w}^t)-F_k^*\right)\right]+\left(F(\mathbf{w}^t)-F^*\right).\end{aligned} \tag{A.21}$$

Plugging Eq. (A.21) into Eq. (A.20), we have:

$$\begin{aligned}A &= \gamma\sum_{k\in S^t}\left[\frac{1}{2\alpha}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2+\alpha L\left(F_k(\mathbf{w}^t)-F_k^*\right)\right]\\ &\quad -\gamma\left(F(\mathbf{w}^t)-F^*\right)+4L\alpha^2\Gamma\\ &= \gamma(\alpha L-1)\sum_{k\in S^t}\left(F_k(\mathbf{w}^t)-F^*\right)+\left(4L\alpha^2\Gamma+\gamma\alpha L\right)\Gamma+\frac{\gamma}{2\alpha}\sum_{k\in S^t}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2.\end{aligned} \tag{A.22}$$

We have:

$$\alpha L-1\leqslant -\frac{3}{4}\leqslant 0, \tag{A.23}$$

$$\sum_{k\in S^t}\left(F_k(\mathbf{w}^t)-F^*\right)=F(\mathbf{w}^t)-F^*\geqslant 0, \tag{A.24}$$

$$\Gamma\geqslant 0, \tag{A.25}$$

$$4L\alpha^2+\gamma\alpha L\leqslant 6\alpha^2 L, \tag{A.26}$$

Therefore:

$$A\leqslant 6L\alpha^2\Gamma+\sum_{k\in S^t}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2. \tag{A.27}$$

Plugging Eq. (A.27) and Eq. (A.8) into Eq. (A.19), we have:

$$\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2\leqslant(1-\mu\alpha)\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2+2\sum_{k\in S^t}\left\|\mathbf{w}^t-\mathbf{w}_k^t\right\|^2+6\alpha^2L\Gamma. \tag{A.28}$$

Theorem 1. We define:

$$\Delta_t=\mathbb{E}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2, \tag{A.29}$$

By Lemma 2, we have:

$$\Delta_{t+1}\leqslant(1-\alpha\mu)\Delta_t+\alpha^2B, \tag{A.30}$$

where $B=6L\Gamma+2(t-t_0)^2G^2$.

We assume that α is a diminishing stepsize, and $\alpha_t=\frac{\beta}{t+\gamma}$, where $\beta>\frac{1}{\mu}$ and $\gamma>0$.

The global model has the following Convergence:

$$\Delta_t\leqslant\frac{v}{\gamma+t}, \tag{A.31}$$

where

$$v=max\left\{\frac{\beta^2B}{\beta\mu-1},(\gamma+1)\Delta_1\right\}. \tag{A.32}$$

Proof 3. We prove it by induction.

First, the definition of v can ensures that when $t=1,\Delta_1\leqslant\frac{v}{\gamma+1}$.

We assume that for some $t,\Delta_t\leqslant\frac{v}{\gamma+t}$.

We prove that for $t+1,\Delta_t+1\leqslant\frac{v}{\gamma+t+1}$:

$$\begin{aligned}\Delta_{t+1} &\leqslant (1-\alpha_t\mu)\Delta_t + \alpha_t^2 B \\ &\leqslant \left(1-\frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2} + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}v\right] \\ &\leqslant \frac{v}{t+\gamma+1}\end{aligned} \tag{A.33}$$

References

[1] J. Konecný, H.B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence, CoRR abs/1610.02527 (2016). arXiv:1610.02527. URL:http://arxiv.org/abs/1610.02527.

[2] J. Konecný, Stochastic, distributed and federated optimization for machine learning, CoRR abs/1707.01155 (2017). arXiv:1707.01155. URL:http://arxiv.org/abs/1707.01155.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, Fort Lauderdale, USA, Vol. 54, PMLR, 2017, pp. 1273–1282.

[4] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, Federated learning with non-iid data, CoRR abs/1806.00582 (2018).

[5] W. Huang, T. Li, D. Wang, S. Du, J. Zhang, T. Huang, Fairness and accuracy in horizontal federated learning, Inf. Sci. 589 (2022) 170–185.

[6] T. Li, M. Sanjabi, A. Beirami, V. Smith, Fair resource allocation in federated learning, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. URL:https://openreview.net/forum?id=ByexElSYDr.

[7] J. Mo, J.C. Walrand, Fair end-to-end window-based congestion control, IEEE/ACM Trans. Netw. 8 (5) (2000) 556–567.

[8] T. Lan, D.T.H. Kao, M. Chiang, A. Sabharwal, An axiomatic theory of fairness in network resource allocation, in: INFOCOM 2010. 29th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 15–19 March 2010, San Diego, CA, USA, IEEE, 2010, pp. 1343–1351.

[9] T. Li, A. Beirami, M. Sanjabi, V. Smith, Tilted empirical risk minimization, in: in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event May 3–7, 2021, OpenReview.net, Austria, 2021.

[10] R.C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: A client level perspective, CoRR abs/1712.07557 (2017). arXiv:1712.07557. URL: http://arxiv.org/abs/1712.07557.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R.S. Zemel, Fairness through awareness, in: S. Goldwasser (Ed.), Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, ACM, 2012, pp. 214–226.

[12] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 3315–3323.

[13] M.B. Zafar, I. Valera, M. Gomez-Rodriguez, K.P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3–7, 2017, ACM, 2017, pp. 1171–1180.

[14] Z. Wang, X. Fan, J. Qi, C. Wen, C. Wang, R. Yu, Federated learning with fair averaging, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event/ Montreal, Canada, 19–27 August 2021, ijcai.org, 2021, pp. 1615–1623.

[15] M. Mohri, G. Sivek, A.T. Suresh, Agnostic federated learning, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 4615–4625. URL:http://proceedings.mlr.press/v97/mohri19a.html.

[16] L. Lyu, J. Yu, K. Nandakumar, Y. Li, X. Ma, J. Jin, H. Yu, K.S. Ng, Towards fair and privacy-preserving federated deep models, IEEE Trans. Parallel Distributed Syst. 31 (11) (2020) 2524–2541.

[17] X. Xu, L. Lyu, Towards building a robust and fair federated learning system, CoRR abs/2011.10464 (2020). arXiv:2011.10464. URL:https://arxiv.org/abs/2011.10464.

[18] J. Zhang, C. Li, A. Robles-Kelly, M.S. Kankanhalli, Hierarchically fair federated learning, CoRR abs/2004.10386 (2020). arXiv:2004.10386. URL:https://arxiv.org/abs/2004.10386.

[19] W. Du, D. Xu, X. Wu, H. Tong, Fairness-aware agnostic federated learning, CoRR abs/2010.05057 (2020). arXiv:2010.05057. URL:https://arxiv.org/abs/2010.05057.

[20] F.A. Farris, The gini index and measures of inequality, Am. Math. Mon. 117 (10) (2010) 851–864, URL:http://www.jstor.org/stable/10.4169/000298910X523344.

[21] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: I.S. Dhillon, D.S. Papailiopoulos, V. Sze (Eds.), Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2–4, 2020, mlsys.org, 2020. URL:https://proceedings.mlsys.org/book/316.pdf.

[22] C.T. Dinh, N.H. Tran, T.D. Nguyen, Personalized federated learning with moreau envelopes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, 2020.

[23] T. Sahni, C. Chandak, N.R. Chedeti, M. Singh, Efficient twitter sentiment classification using subjective distant supervision, in: 9th International Conference on Communication Systems and Networks, COMSNETS 2017, Bengaluru, India, January 4–8, 2017, IEEE, 2017, pp. 548–553.

[24] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543.

[25] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Tech. rep., Citeseer (2009).

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States., 2012, pp. 1106–1114.