FBLG: A Local Graph Based Approach for Handling Dual Skewed Non-IID Data in Federated Learning

Yi Xu [1] , **Ying Li** [1] , **Haoyu Luo** [2] , **Xiaoliang Fan** [3] and **Xiao Liu**[4]
[1]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2]College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China
[3]Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China
[4]School of Information Technology, Deakin University, Geelong, Australia
xuyi1023@126.com, yingli@stu.ahu.edu.cn, haoyuluo@scau.edu.cn, fanxiaoliang@xmu.edu.cn, xiao.liu@deakin.edu.au

Abstract

In real-world situations, federated learning often needs to process non-IID (non-independent and identically distributed) data with multiple skews, causing inadequate model performance. Existing federated learning methods mainly focus on addressing the problem with a single skew of non-IID, and hence the performance of global models can be degraded when faced with dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients. To address the problem with dual skewed non-IID data, in this paper, we propose a federated learning algorithm based on local graph, named *FBLG*. Specifically, to address the label distribution skew, we firstly construct a local graph based on clients' local losses and Jensen-Shannon (JS) divergence, so that similar clients can be selected for aggregation to ensure a highly consistent global model. Afterwards, to address the sample size skew, we design the objective function to favor clients with more samples as models trained with more samples tend to carry more useful information. Experiments on four datasets with dual skewed non-IID data demonstrate *FBLG* outperforms nine baseline methods and achieves up to 9% improvement in accuracy. Simultaneously, both theoretical analysis and experiments show *FBLG* can converge quickly.

1 Introduction

Federated learning has been widely applied in many fields. For example, in the medical field, federated learning can help medical institutions of different levels share patient data to provide more accurate treatment plans. In the Internet of Things (IoT), federated learning is used for shared training among multiple devices to improve the performance of smart devices. Additionally, federated learning has been applied in various fields such as finance and transportation to address the data privacy and distributed learning [Li *et al.*, 2020a]. With the federated learning framework, clients can collaboratively train models without exposing data. Firstly, clients train models on local devices. Subsequently, the clients send updates of local models to the central server, which aggregates these updated local models and sends the resulting global model to each client [McMahan *et al.*, 2017].

A major challenge in federated learning is the non-IID data, which arises when the data distributions on different devices are not independent and identically distributed [Liao *et al.*, 2023; Shang *et al.*, 2022]. The non-IID data can cause drift in the optimization of both local and global models, resulting in slower convergence [Karimireddy *et al.*, 2020; Li *et al.*, 2020c]. Existing study indicates that causes of non-IID data can be subdivided into five categories: feature distribution skew, label distribution skew, concept drift with different features, concept drift with different labels, and sample size skew [Kairouz *et al.*, 2021]. However, nearly all existing federated learning methods focus on researching non-IID data with only one specific type of skew. For instance, FedLC [Zhang *et al.*, 2022] proposed a fine-grained calibrated cross-entropy loss to reduce the bias in local updates to improve the performance of global models with label distribution skews; Tijani *et al.*[2021] proposed a data extension strategy aimed at generating placeholders for absent classes within a local dataset to address the label distribution skew.

However, in real-world situations, lots of non-IID data comprises dual or even multiple skews. For example, consider three hospitals: a tertiary hospital, a children's hospital, and a tumor hospital. The distribution of tumor labels among these hospitals tends to be skewed, and simultaneously, the sample sizes also tend to be skewed due to differing sizes of these hospitals [Wu *et al.*, 2023]. Existing studies [Hsu *et al.*, 2019; Hsieh *et al.*, 2020] have demonstrated that the sole skew caused by heterogeneous label distributions among clients can reduce the performance of the global model by 40%. Furthermore, the presence of dual skews caused by heterogeneous label distributions and sample sizes among clients can lead to performance degradation by 56% for the global model, highlighting the substantial impact of dual skews on global model performance. As the complexity of addressing non-IID data consisting of multiple skews will be significantly increased, it remains an open problem that is far from being resolved [Li *et al.*, 2022]. So far, there are very limited studies focusing on non-IID data characterized by dual or multiple skews [Zhu *et al.*, 2021]. Therefore, as an initial effort, our primary focus is to address the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients. We hope this work will provide a solid foundation for handling more intricate cases with multiple skews which require in-depth research in the future.

In this paper, we propose a Federated learning algorithm Based on Local Graph, named *FBLG* that can simultaneously address the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients. The main components of the algorithm include:

(i) Addressing the skew caused by heterogeneous label distributions among clients. After each communication round, the server firstly sorts clients according to their local losses. Then, the server selects clients with larger local losses to construct a local graph and calculates the JS divergence among clients as the weights of edges in the local graph. Based on the local graph, clients with larger local losses and higher similarities are selected for aggregation to make the global model highly consistent.

(ii) Addressing the skew caused by heterogeneous sample sizes among clients. Considering that the model trained with more samples tends to carry more useful information, the sample size is used to further select clients with more samples when designing the objective function.

In summary, the algorithm considers selecting clients with larger local losses, higher similarities, and more samples for aggregation. Experimental results demonstrate the *FBLG* algorithm can achieve higher accuracy than existing baseline methods on four datasets when both the label distribution and sample size are skewed among clients. Theoretical analysis and experiments show the *FBLG* algorithm can converge quickly.

Our contributions in this paper are as follows:

- We propose the *FBLG* algorithm to address dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients.
- We construct a local graph based on clients' local losses and JS divergence among clients. Subsequently, similar clients are selected for aggregation based on the local graph to address the label distribution skew. Additionally, considering that models trained with more samples tend to contain more useful information, we use the sample size to select clients with more samples when designing the objective function to address the sample size skew.
- Through comparison experiments with nine existing baseline methods and theoretical analysis, it is verified that the *FBLG* algorithm can achieve higher accuracy and quicker convergence under situations with dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients.

2 Related Work

Federated learning is a distributed machine learning method that involves sharing data across multiple clients for model training while protecting individual privacy. Existing federated learning frameworks are divided into vertical federated learning and horizontal federated learning [Zhang *et al.*, 2021a]. This paper is based on the horizontal federated learning framework, where participants typically share identical features while possessing distinct sample sets.

However, due to the difference in distributions of data owned by various clients participating in federated learning, non-IID data has become an important issue. The existence of non-IID data can lead to performance degradation in models, thus affecting the overall effectiveness of federated learning [Zhao *et al.*, 2018; Li *et al.*, 2022].

In recent years, some progress has been made to address non-IID data. However, current work mainly focuses on solving non-IID data with only one certain kind of skew among clients, such as label distribution skew, feature distribution skew, or sample size skew. To address the label distribution skew, Ramakrishna *et al.* [2022] proposed approximate inference methods for category label distribution based on parameter updates of clients; FedOV [Diao *et al.*, 2023] deleted the original features of a few classes and learned them as adversaries. To address the feature distribution skew, FedBN [Li *et al.*, 2021b] added a batch normalization layer to the local model to mitigate feature bias before model aggregation; FedRDN [Yan and Zhu, 2023] randomly injected statistical information from the entire federation's dataset into clients' data. To address the sample size skew, Wang *et al.* [2021] monitored and designed a new loss to make weight updates proportional to the number of samples in different categories; Zhang *et al.* [2021b] proposed a client selection system, enabling clients to decide participation in each training round based on their individual and global data distribution probabilities.

However, existing studies [Hsu *et al.*, 2019; Hsieh *et al.*, 2020] have shown that dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients degrade the performance more than the sole skew caused by heterogeneous label distributions among clients. Therefore, based on existing studies, this paper addresses the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients to improve the effectiveness of federated learning.

3 Our Method

3.1 Preliminaries

In our work, we suppose there are N clients. For the k-th client, its sample set is D_k, the number of samples is n_k, and the global model delivered by the server at t-th round is θ^t. Firstly, for $\forall d \in D_k$, let $f_d(\theta^t)$ be the local loss of each sample on the client, then the local loss for the k-th client is denoted as

$$F_k(\theta^t) = \frac{1}{n_k} \sum_{d \in D_k} f_d(\theta^t) \tag{1}$$

For the k-th client, the optimization objective is to find the local model θ_k^* that minimizes the loss, denoted as

$$\theta_k^* = \arg\min_{\theta_k} \{F_k(\theta_k)\} \tag{2}$$

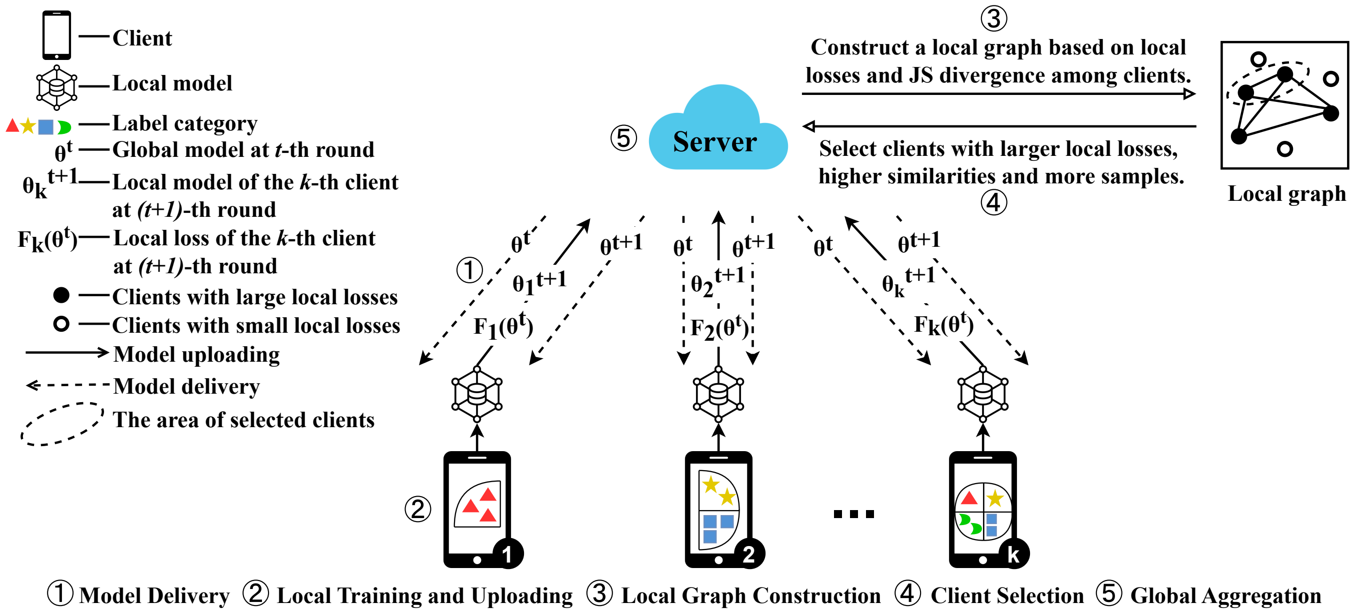


Figure 1: The framework of *FBLG*.

We use the stochastic gradient descent (SGD) to optimize the local loss $F_k(\theta^t)$ on the k-th client to obtain the local model θ_k^{t+1} after a new round of local training. Then, we upload $F_k(\theta^t)$ and θ_k^{t+1} at the same time to the server, which constructs the local graph. Based on the local graph, we select clients to achieve aggregation.

3.2 Framework of Our Proposed *FBLG*

This section introduces the framework of our proposed *FBLG*, which is shown in Fig. 1 and Algorithm 1. Assuming that the global model θ^t has been obtained after completing the t-th round of iteration, the $(t+1)$-th round in *FBLG* includes the following steps. Each step is labeled with their corresponding line numbers in Algorithm 1.

Step 1: The server delivers the global model θ^t (Line 3).

Step 2: Each client performs local training based on θ^t. It obtains the local loss $F_k(\theta^t)$ and the trained local model θ_k^{t+1}, then uploads them to the server (Lines 4 to 7).

Step 3: The server constructs a local graph based on the local loss $F_k(\theta^t)$ and the JS divergence among clients (Lines 9 to 10).

Step 4: The server uses the local graph to select clients with larger local losses, higher similarities, and more samples (Lines 11 to 12).

Step 5: The server aggregates selected clients and generates the global model θ^{t+1} (Lines 13 to 16).

The framework of *FBLG* is consistent with most federated learning methods in stages of global model delivery, client local training, and model aggregation, while the key difference lies in the selection of clients (*i.e.* Steps 3 and 4).

3.3 Local Graph Construction

Next, we will introduce the construction of the local graph (*i.e.* Step 3). Firstly, we need to select the top $M = C \times N$

Algorithm 1: *FBLG* algorithm

Input: The global model θ, the sizes of clients' local data n, the total round T, the client number N, the number of local updating steps E, the learning rate η, and the proportion C

1 Initialize the global model θ^0 and $M = C \times N$
  foreach *communication round* $t = 1, 2, ..., T$ **do**
2   **foreach** *client* $k = 1, 2, ..., N$ **do**
3   The server delivers the global model θ^t to the client k
4   **foreach** *local updating step* $u = 1, 2, ..., E$ **do**
5         $\theta_{k,u+1}^{t+1} \leftarrow \theta_{k,u}^{t} - \eta \nabla F_k(\theta_{k,u}^{t})$
6   **end**
7   The client k uploads $F_k(\theta^t)$ and θ_k^{t+1} to the server
8   **end**
9   The server selects the top M clients with the maximum $F_k(\theta^t)$ as the candidate set c_{t+1}
10   Create the local graph G for the candidate set c_{t+1} based on the techniques in Sec.3.3
11   Compute the shortest-path distance of each pair nodes on G by Floyd Algorithm to obtain H
12   The server selects clients as s_{t+1} by
$$\max_{s_{t+1} \leq c_{t+1}} \left(\frac{s_{t+1}^{\top} H s_{t+1}}{M(M-1)} + \frac{s_{t+1}^{\top} n_k s_{t+1}}{\sum_{k \in s_{t+1}} n_k}\right)$$
13   **foreach** *client* $k \in s_{t+1}$ **do**
14      $w = \frac{n_k}{\sum_{k \in s_{t+1}} n_k}$
15   The server aggregates received local models $\theta^{t+1} = \sum_{k \in s_{t+1}} w \theta_k^{t+1}$
16   **end**
17 **end**

clients with larger local losses from N clients, where C is the proportion of clients with larger local losses selected, and we denote these M clients as $c_1, c_2, \cdots, c_M$. Then, we use clients $c_1, c_2, \cdots, c_M$ to construct a local graph and denote the local graph as $G(\gamma, V)$, where $\gamma = \{c_1, c_2, \cdots, c_M\}$ is the node set of the local graph $G(\gamma, V)$, $V \in R^{M \times M}$ is the adjacency matrix of the local graph $G(\gamma, V)$, and the element V_{ij} is the weight between any two clients i and j, which is mainly used to characterize the similarity between any two clients i and j.

Existing similarity methods usually employ the cosine distance, which only focuses on the direction of vectors, disregarding their specific magnitudes. This can lead to poor results when computing the similarity of non-sparse data. Therefore, we use JS divergence to measure the similarity between clients. In addition, once the feature vectors of clients are given, we can easily get the adjacency matrix among clients. However, it is crucial to note that feature vectors may leak clients' sensitive information. Taking inspiration from FedGS [Wang *et al.*, 2023], we introduce Gaussian noise in the computation of JS divergence. The computation of JS divergence is denoted as

$$S_{ij} = \frac{1}{2} \int e_i \log \frac{e_i}{\frac{e_i + e_j}{2}} dx + \frac{1}{2} \int e_j \log \frac{e_j}{\frac{e_i + e_j}{2}} dx \tag{3}$$

where $e_i = \theta_i(\varepsilon)[r]$ is the average of the r-th layer's network embedding on a batch for each client after we feed the batch of random Gaussian noise $\varepsilon \sim N(\mu, \Sigma)$ to all the locally trained models, while μ and Σ are respectively the mean and covariance of a small validation dataset owned by the server.

Based on the JS divergence S_{ij} between any two clients i and j, the weight V_{ij} between any two clients i and j is denoted as

$$V_{ij} = \frac{1}{S_{ij} + 1} \tag{4}$$

Based on the local graph $G(\gamma, V)$, the following explains how to select clients.

3.4 Client Selection

Next, we will introduce how to select clients with larger local losses, higher similarities, and more samples based on the local graph (*i.e.* Step 4). To achieve quicker error convergence of the model [Cho *et al.*, 2020], we naturally prioritize the selection of clients with larger local losses, as we consider clients with larger local losses as one of our construction indicators when constructing the local graph.

Then, to address the label distribution skew among clients, we select clients with higher similarities for aggregation. Firstly, we calculate the shortest path distance matrix $H = [h_{ij}]_{M \times M}$ between each node pair in the local graph $G(\gamma, V)$ by the Floyd algorithm, where h_{ij} is the shortest path distance between any two clients i and j. Subsequently, based on the shortest path distance matrix H, we use an optimization objective to select a larger shortest path. The optimization objective ($O1$) is denoted as

$$\max_{s_{t+1} \le c_{t+1}} \left(\frac{s_{t+1}^\top H s_{t+1}}{M(M-1)} \right) \tag{5}$$

where $s_{t+1} = \left\{ s_{t+1}^1, s_{t+1}^2, \cdots, s_{t+1}^M \right\}$ is the selection result at $(t+1)$-th round, $s_{t+1}^k \in (0, 1)$, $1 \le k \le M$, when $s_{t+1}^k = 1$, it means that the k-th client is selected to participate in $(t+1)$-th round of aggregation, conversely, when $s_{t+1}^k = 0$, it means that the k-th client is not selected, M is the number of nodes in the local graph $G(\gamma, V)$, and $M(M-1)$ is the number of node pairs in the local graph $G(\gamma, V)$.

According to the optimization objective of Eq. (5) and the definition of Eq. (4), we can observe that when the shortest path distance matrix H is relatively large, it implies that the JS divergence between clients needs to be as small as possible and hence we need to select clients with higher similarities.

Finally, to address the sample size skew among clients, we prioritize selecting clients with more samples, as models trained with more samples tend to carry more useful information. To this end, the optimization objective ($O2$) is denoted as

$$\max_{s_{t+1} \le c_{t+1}} \left(\frac{s_{t+1}^\top n_k s_{t+1}}{\sum_{k \in s_{t+1}} n_k} \right) \tag{6}$$

where $\sum_{k \in s_{t+1}} n_k$ is the total sample size of all clients.

In summary, to simultaneously address the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients, we define the total optimization objective ($O1 + O2$) when selecting clients as

$$\max_{s_{t+1} \le c_{t+1}} \left(\frac{s_{t+1}^\top H s_{t+1}}{M(M-1)} + \frac{s_{t+1}^\top n_k s_{t+1}}{\sum_{k \in s_{t+1}} n_k} \right) \tag{7}$$

3.5 Global Aggregation

After selecting clients based on the local graph through Eq. (7), local models are aggregated using the traditional weighted aggregation method to produce the global model (*i.e.* Step 5). During the aggregation process, considering that clients with more samples should have a larger weight in the aggregation, we define the weights based on sample sizes within clients as

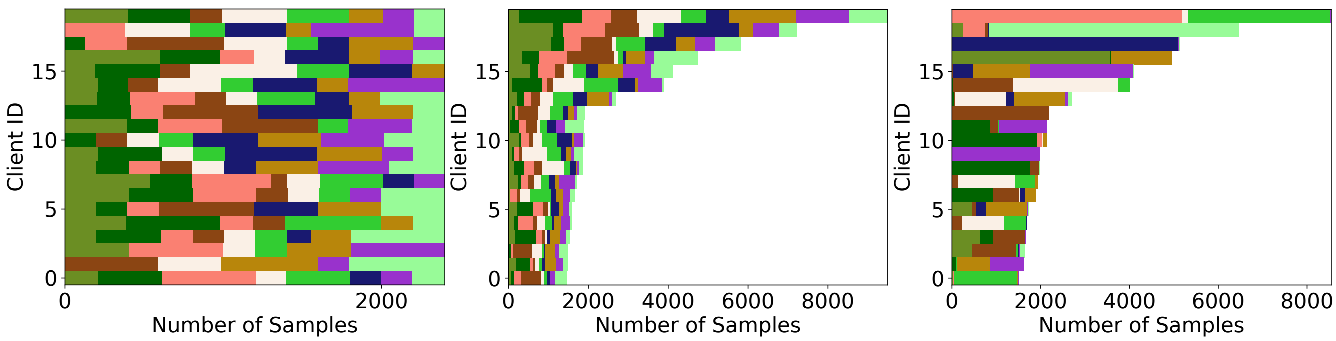
$$w = \frac{n_k}{\sum_{k \in s_{t+1}} n_k} \tag{8}$$

This approach allows us to further alleviate the sample size skew among clients by assigning larger weights to clients with larger sample sizes and smaller weights to clients with smaller sample sizes. Finally, the global model at $(t+1)$-th round is denoted as

$$\theta^{t+1} = \sum\nolimits_{k \in s_{t+1}} w \theta_k^{t+1} \tag{9}$$

3.6 Convergence Analysis

We analyze the convergence of our proposed *FBLG* from a global perspective under the assumption that the loss function is non-convex. First, according to Bubeck *et al.* [2015], if the loss function $\nabla F(\theta)$ is β-smooth, then for any θ, $\theta' \in \mathbb{R}^d$, there exists $\|\nabla F(\theta) - \nabla F(\theta')\| \le \beta \|\theta - \theta'\|$. Second, according to Tian *et al.* [2022], if $F(\theta)$ is locally convex, then for any θ, $\theta' \in \mathbb{R}^d$, $\varphi \in [0, 1]$, the distance between θ



(a) (b) (c)

Figure 2: Visualization of three skewed data.

and the locally optimum θ' within a radius r(>0), there exists $F\left(\varphi\theta + (1-\varphi)\theta'\right) \leq \varphi F(\theta) + (1-\varphi)F(\theta')$.

Next, we first prove that the model θ based on our proposed *FBLG* converges within the range selected by clients each time during the training process.

Theorem 1. *If $\eta < 1/\beta$, β is a constant, then there exists $\left\|F(\theta^{t+1}) - F(\theta^*)\right\| \leq \|F(\theta^t) - F(\theta^*)\|$, where $F(\theta^t)$ represents the loss function of the global model at t-th round, θ^t and θ^* represents the model and the optimized model defined in Eq.* (2) *on the server, respectively.*

Proof. See Appendix A. □

We then discuss the reasons why convergence improves when similar clients are aggregated. In federated learning, there are N clients with sample sets $D_1, D_2, ...D_n$, each of which belongs to one of the $p^l(l = 1, 2, ..., (\leq n))$ distributions. Assuming that the stochastic gradient $g^l(\cdot)$ obtained from the distribution p^l at t-th round is unbiased, *i.e.* $\mathbb{E}[g^l(\theta^t)] = \nabla F^l(\theta^t)$. Since clients selected based on the *FBLG* are highly similar, it is natural to assume that data of selected clients come from almost the same distribution.

Theorem 2. *Suppose that FBLG selects a set of local models trained with the same distribution of datasets. Compared with the FedAvg, we get $\mathbb{E}\|\theta_l^t - \theta_l^*\|^2 \leq \mathbb{E}\left\|\bar{\theta}^t - \theta_l^*\right\|^2$, where θ_l^* is the optimized model for the dataset fitting the distribution p^l, θ_l^t denotes the global model aggregated by FBLG for the distribution p^l, while $\bar{\theta}^t$ denotes the uniform global model of FedAvg at t-th round.*

Proof. See Appendix B. □

We use the loss function to measure convergence and theoretically justify our proposed *FBLG*. If the loss function value converges stably to 0, it equivalently reflects that the trained model can converge to the optimal.

4 Evaluation

This section aims to answer the following research questions:

RQ1. How do the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients affect the accuracy of federated learning?

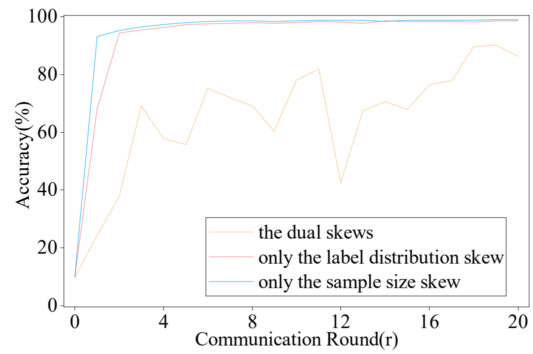


Figure 3: Impact of the dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients.

RQ2. How is the performance of our proposed *FBLG* method compared with baseline methods?

RQ3. How do different components (including similarity metrics and objective functions) affect our proposed *FBLG*?

4.1 Experimental Settings

Datasets. According to most papers in federated learning [McMahan *et al.*, 2017; Wang *et al.*, 2020a; Huang *et al.*, 2022], we validate our proposed *FBLG* algorithm on four commonly used datasets: MNIST [LeCun *et al.*, 1998], Fashion-MNIST (FMNIST) [Xiao *et al.*, 2017], CIFAR10 [Krizhevsky *et al.*, 2009] and SVHN [Netzer *et al.*, 2011]. To create data with dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients, we assign label distribution and sample size to 20 clients through Dirichlet distribution *i.e.* $Y_k \sim Dir(\alpha\Upsilon)$, where α denotes the degree of skews among clients, Υ denotes the global label distribution [Hsu *et al.*, 2019]. The smaller α indicates the data is more skewed. The sole label skew is set by shards [Wang *et al.*, 2023]. Our codes, some supplementary experiments and appendices mentioned in the paper are available at https://github.com/YingLi-Y/FBLG.git.

Baselines. We compare our *FBLG* method with: (i) FedAvg [McMahan *et al.*, 2017] based on weight aggregation, (ii) MDSample [Li *et al.*, 2020b] based on the client's local

Dataset	MNIST			CIFAR10			FMNIST			SVHN		
Method / α	0.8	0.05	0.01	0.8	0.05	0.01	0.8	0.05	0.01	0.8	0.05	0.01
FedAvg	98.86	97.99	97.32	60.28	42.60	41.57	89.04	81.97	77.26	91.09	78.18	61.74
MDSample	98.80	98.07	97.41	60.77	40.78	47.49	89.12	82.63	80.35	91.20	69.32	62.27
Power-Of-Choice	98.83	**98.36**	**98.40**	58.16	34.31	38.03	87.89	83.24	81.11	92.17	71.58	60.82
FedProx (ϕ = 0.1)	98.63	96.77	95.96	60.99	45.75	45.86	**89.23**	79.46	77.32	92.40	73.12	59.76
Moon	98.31	92.77	90.54	58.32	37.59	38.13	88.39	80.48	68.83	90.13	67.76	56.87
Scaffold	98.80	97.08	92.89	60.76	37.27	39.12	88.62	77.42	67.47	91.87	70.08	60.45
FedAvgM	98.77	97.66	97.30	61.00	45.83	41.56	88.99	81.90	79.66	91.50	78.54	66.77
FedNova	98.60	95.49	96.93	57.87	32.65	39.59	88.70	77.48	75.44	91.47	70.07	59.46
FedGS	98.62	97.98	97.45	60.08	46.49	38.03	87.97	82.77	82.23	91.94	80.35	68.84
FBLG (C = 0.5)	98.86	**98.36**	98.27	61.08	**54.01**	**55.08**	**89.23**	**86.34**	**83.54**	**92.76**	**85.38**	**78.10**
FBLG (C = 0.4)	**98.93**	98.24	98.20	**61.55**	52.51	53.30	88.51	86.03	83.22	92.32	84.79	71.47

Table 1: Accuracy (%) comparison results on four datasets under different degrees of skews. The best results are marked in bold. The second-best results are underlined.

data size, (iii) Power-Of-Choice [Cho *et al.*, 2020] based on the client's local loss, (iv) FedProx [Li *et al.*, 2020b] based on the proximal term, (v) Moon [Li *et al.*, 2021a] based on the model comparison loss, (vi) Scaffold [Karimireddy *et al.*, 2020] based on control variables, (vii) FedAvgM [Hsu *et al.*, 2019] based on regularization, (viii) FedNova [Wang *et al.*, 2020b] based on the speed of local training on the client, (ix) FedGS [Wang *et al.*, 2023] based on the data distribution dependency graph and the sampling frequency of the client.

Parameter Settings. For each dataset, the number of samples selected by the client for one training session is $B = 64$, the number of local iterations is $E = 5$, the learning rate is $\eta = 0.05$, the learning rate is fully attenuated by a factor of 0.998, and the optimization algorithm used for local training of clients is SGD.

Implementation. All our experiments are run on the AIStation server with 1 NVIDIA A100-SXM4-40GB and 4 CPUs. All codes are implemented in Pytorch 1.12.1.

4.2 Impact of Dual Skewed Non-IID Data (RQ1)

We take the MNIST dataset as an example to more intuitively illustrate the label distribution skew among clients, the sample size skew among clients, and the dual skews caused by heterogeneous label distributions and sample sizes among clients. The visualization of the three types of data mentioned above is shown in Fig. 2, where the x-axis denotes the number of samples, the y-axis denotes the client ID, and the color of the block denotes the label type of samples.

- Fig. (a) visualizes the case where only the label distribution is skewed among clients when the number of shards is 12. Here, each client has a consistent sample size, but the label distribution of each client is inconsistent.
- Fig. (b) visualizes the case where only the sample size is skewed among clients when α = 17. Here, each client has a consistent label distribution but the sample size of each client is inconsistent.
- Fig. (c) visualizes the case where the label distribution and sample size are both skewed among clients when α = 0.8. Here, some clients have 3 labels, while others have only 1 label, and the sample size per client is inconsistent.

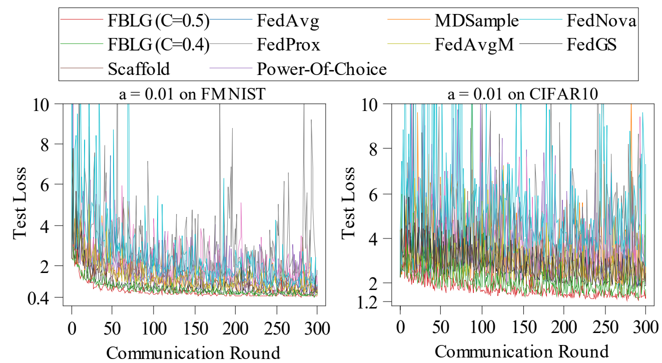


Figure 4: Test loss curves respectively on FMNIST and CIFAR10 when α = 0.01.

To demonstrate the impact of data with only the label distribution skew among clients, only the sample size skew among clients, and the dual skews caused by heterogeneous label distributions and sample sizes among clients in federated learning, we use the classic FedAvg algorithm as an example. We plot the classification accuracies of the three skews with 20 clients over 20 communication rounds with the line graph, as shown in Fig. 3. We observe that only the label distribution skew among clients and only the sample size skew among clients can achieve relatively high accuracy within the first 4 communication rounds. Specifically, only the sample size skew among clients can even achieve 99.02% accuracy within 20 communication rounds. However, the dual skews caused by heterogeneous label distributions and sample sizes among clients can only reach 90.20% accuracy within 20 communication rounds and the convergence of the global model is unstable. It is clear that dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients has a greater impact on the performance of the global model.

Similarity Metrics	MNIST	CIFAR10	FMNIST	SVHN	AVG
JS divergence	**98.86**	**61.08**	**89.23**	**92.76**	**85.48**
cosine distance	98.79	60.65	89.17	92.12	85.18
euclidean distance	98.84	60.70	89.02	92.30	85.21

Table 2: Impact of similarity metrics when $\alpha = 0.8$. The best results are marked in bold.

Objective Functions	MNIST	CIFAR10	FMNIST	SVHN
FBLG ($O1$)	98.60	58.53	88.57	92.32
FBLG ($O2$)	98.79	60.54	89.14	92.40
FBLG ($O1 + O2$)	**98.86**	**61.08**	**89.23**	**92.76**

Table 3: Impact of objective functions when $\alpha = 0.8$. The best results are marked in bold.

4.3 Results of Performance Comparison (RQ2)

We conducted experiments on four datasets with 20 clients over 300 communication rounds when α = {0.8, 0.05, 0.01}, and the accuracy results are shown in Table 1. We observe that: (i) the accuracy of most baseline methods decreases as α decreases, which indicates that highly skewed data can seriously affect the performance of baseline methods. (ii) our proposed *FBLG* (C = 0.5) can achieve high accuracy on almost all four datasets, where C is the proportion of clients with larger local losses selected. When α = 0.05, it can improve at least 6.02% on the CIFAR10 dataset, and even achieve relatively high accuracy on extremely skewed data (*i.e.* α = 0.01), while existing nine baseline methods cannot. In addition, our proposed *FBLG* can achieve higher classification accuracy on MNIST and CIFAR10 datasets at C = 0.4 when α = 0.8. (iii) Although Power-Of-Choice and FedProx (ϕ = 0.1) can achieve the same high accuracy as *FBLG* (C = 0.5) proposed in this paper on two of the experimental results, Power-Of-Choice outperforms ours in an experimental result, we plot the test loss curves of each algorithm under extremely skewed data (*i.e.* α = 0.01) respectively on the FMNIST and CIFAR10 datasets, as shown in Fig. 4. We observe that the test loss of *FBLG* is small and converges quickly, consistently outperforming other baseline methods. Thus, our proposed *FBLG* can effectively address the impact of dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients. More details about the data visualization and test loss are in Appendix C.

4.4 Ablation Study (RQ3)

Similarity Metrics. We first verify the impact of the similarity metric based on JS divergence adopted in this paper on our *FBLG's* performance when $\alpha = 0.8$. Here, we respectively replace the similarity metric with cosine distance and euclidean distance, then plot the impact of the three similarity metrics on the classification accuracy of our proposed *FBLG* into a table, as shown in Table 2. We observe that our proposed *FBLG* performs best with the similarity metric based on JS divergence, achieving up to 0.46% higher accuracy on the SVHN dataset. This preference is attributed to the robust nature of JS divergence in handling extreme values or outliers, making it more capable of addressing skewed data.

Objective Functions. We then compare the objective function that considers both the similarity and the number of samples among clients, the objective function that only considers the similarity among clients, and the objective function that only considers the number of samples among clients, and verify the impact of these three situations on the classification accuracy of our proposed *FBLG* algorithm when $\alpha = 0.8$. We denote Eq. (5) that only considers the similarity among clients to address the label distribution skew among clients as *FBLG (O1)* and denote Eq. (6) that only considers the number of samples among clients to address the sample size skew among clients as *FBLG (O2)*. Simultaneously, the Eq. (7) that considers both the similarity and the number of samples among clients to address dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients is denoted as *FBLG (O1+O2)*. The impact of the above three situations on the classification accuracy of our proposed *FBLG* is drawn into a table, as shown in Table 3. We observe that considering both the similarity and the number of samples among clients in Eq. (7) can bring better performance than only considering the similarity among clients or only considering the number of samples among clients.

5 Conclusion

In this paper, we proposed a new federated learning algorithm based on local graph (*FBLG*) to address dual skewed non-IID data caused by heterogeneous label distributions and sample sizes among clients. Specifically, (i) To address the label distribution skew, we construct a local graph based on the local losses of clients and the JS divergence among clients. Based on the local graph, similar clients are selected for aggregation to make the global model highly consistent; (ii) To address the sample size skew, we use the sample size to select clients with more samples when designing the objective function. Experimental results demonstrated the accuracy of our proposed *FBLG* is higher than that of baseline methods. Especially with the increasing degrees of data skewness, the advantage of *FBLG* becomes more obvious. Meanwhile, both theoretical analysis and experimental results have successfully proven that our proposed *FBLG* can converge quickly. In the future, we will explore to address the problem of non-IID data consisting of more complex multiple skews.

Acknowledgements

The research was supported by National Natural Science Foundation of China (No. 62076002, 61402005, 62272403), and Natural Science Foundation of Anhui Province, China (No.2008085MF194).

References

[Bubeck, 2015] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, pages 231–357, 2015.

[Cho *et al.*, 2020] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

[Diao *et al.*, 2023] Yiqun Diao, Qinbin Li, and Bingsheng He. Towards addressing label skews in one-shot federated learning. In *ICLR*, 2023.

[Hsieh *et al.*, 2020] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *ICML*, pages 4387–4398, 2020.

[Hsu *et al.*, 2019] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[Huang *et al.*, 2022] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, pages 10143–10153, 2022.

[Kairouz *et al.*, 2021] 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.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143, 2020.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2020a] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, volume 2, pages 429–450, 2020.

[Li *et al.*, 2020c] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *ICLR*, 2020.

[Li *et al.*, 2021a] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021.

[Li *et al.*, 2021b] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021.

[Li *et al.*, 2022] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *ICDE*, pages 965–978, 2022.

[Liao *et al.*, 2023] Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Huabin Zhu, Yanchao Tan, Jun Wang, and Yue Qi. Hyperfed: hyperbolic prototypes exploration with consistent aggregation for non-iid data in federated learning. In *IJCAI*, pages 3957–3965, 2023.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS*, 2011.

[Ramakrishna and Dán, 2022] Raksha Ramakrishna and György Dán. Inferring class-label distribution in federated learning. In *AISec*, pages 45–56, 2022.

[Shang *et al.*, 2022] Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. In *IJCAI*, pages 2218–2224, 2022.

[Tian *et al.*, 2022] Pu Tian, Weixian Liao, Wei Yu, and Erik Blasch. Wscc: A weight-similarity-based client clustering approach for non-iid federated learning. *IEEE Internet of Things Journal*, 9(20):20243–20256, 2022.

[Tijani *et al.*, 2021] Saheed A. Tijani, Xingjun Ma, Ran Zhang, Frank Jiang, and Robin Doss. Federated learning with extreme label skew: A data extension approach. In *IJCNN*, pages 1–8, 2021.

[Wang *et al.*, 2020a] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *INFOCOM*, pages 1698–1707, 2020.

[Wang *et al.*, 2020b] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[Wang *et al.*, 2021] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *AAAI*, pages 10165–10173, 2021.

[Wang *et al.*, 2023] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Haibing Jin, Peizhen Yang, Siqi Shen, and Cheng Wang. Fedgs: Federated graph-based sampling with arbitrary client availability. In *AAAI*, pages 10271–10278, 2023.

[Wu *et al.*, 2023] Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *IJCAI*, pages 4424–4432, 2023.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Yan and Zhu, 2023] Yunlu Yan and Lei Zhu. A simple data augmentation for feature distribution skewed federated learning. *arXiv preprint arXiv:2306.09363*, 2023.

[Zhang *et al.*, 2021a] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

[Zhang *et al.*, 2021b] Shulai Zhang, Zirui Li, Quan Chen, Wenli Zheng, Jingwen Leng, and Minyi Guo. Dubhe: Towards data unbiasedness with homomorphic encryption in federated learning client selection. In *ICPP*, pages 1–10, 2021.

[Zhang *et al.*, 2022] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *ICML*, pages 26311–26329, 2022.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[Zhu *et al.*, 2021] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.