SUMMARY OF CHANGES

We are immensely pleased to be offered helpful suggestions and have thoroughly revised the manuscript according to the reviewer's comments. In this revision, we have taken the opportunity to address the reviewer's concerns that were kindly drawn to our attention. We thoroughly considered each comment and made changes to clarify the manuscript accordingly. The following items summarize the main changes in the latest revision conducted to the paper.

1) Section I "Introduction" was revised, and a part of the content has been put into section II, Related Work, to improve the manuscript's readability.
2) Section VI "Ablation Study: FedDiskAb" was added to provide a more thorough study of the proposed method and its modules.
3) Section VIII "Problem statement" was provided to better formulate the problem and the goal to achieve.
4) Section VII-D (Discussion) was added to provide a brief discussion about the advantages and drawbacks of the proposed method.
5) Data and source code Github repository link is included for references. The link is attached to the manuscript footnote.
6) Several sections have been revised in terms of writing, including Abstract, Introduction, Related Work, Experiment, and Methodology.
7) Added experiments with a state-of-the-art method, namely FedPCL, for a more comprehensive comparison.

I. RESPONSE TO Associate Editor

-A section of Related Works is needed. -A section of the Problem Statement is needed. -Comparative results with literature state-of-the-art approaches are needed.

Thank you for your recommendation. In this revision, we have addressed all of your concerns. Specifically, we added ”Related Work” section, ”Problem Statement” section and a comparision with a state-of-the-art method, namely FedPCL.

II. RESPONSE TO REVIEWER 1

Comment 1.1 In the abstract, “non-IID data issue” should be concisely discussed and IID must be expanded at its first occurrence.

Response 1.1: Thank you for your suggestion. In this revision, the term “non-IID” has been expanded and briefly discussed in the abstract. We revised the abstract section accordingly as follows.

The change can be listed as follows.

“‘ One of the most challenging issues in federated learning is that the data is often not independent and identically distributed (non-IID). Clients are expected to contribute the same type of data and drawn from one global distribution. However, data are often collected in different ways from different resources. Thus, the data distributions among clients might be different from the underlying global distribution. This creates a weight divergence issue and reduces federated learning performance. This work focuses on improving federated learning performance for skewed data distribution across clients. ...

”””

Comment 1.2 As indicated in Introduction “the primary problem is the divergence of weights, which worsens...”, please discuss this issue with an example and strengthen the motivation of the tackled problem.

Response 1.2: Thank you for your suggestion. We have revised the introduction accordingly and added example to strengen our motivation.

The changes can be seen at page 1, line 21 and also listed as follows.

“‘ .... The primary problem is the divergence of model weights, as found in [2] by Zhao et al. The authors showed that the model’s weights tend to be more diverged for non-IID data compared to that for IDD data. This causes a performance reduction, and it worsens as the data distribution becomes more skewed. For example, the accuracy dropped by about 10% for image dataset Cifar-10 [3], and speech recognition dataset KWS [4] with a non-IID setting ...

”””

Comment 1.3 How the proposed method is effective in resolving the issue of non-IID and privacy leakage in real-time while meeting the weight optimization criteria? Please include this detailed information within the proposed method.

Response 1.3: Thank you for your suggestion. In this version, we have revised the Methodology section accordingly and added an explaination on how the proposed method can help resolve the non-IID and

privacy leakage.

The changes can be seen at page 3, line 7 and also listed as follows.

“‘ .... In this section, a solution is proposed to alleviate the negative impact of distribution skewness across clients for federated learning by adjusting client data distribution during the training process. The proposed method aims to find weights for training samples in order to adjust client data distributions. The remainder of this section introduces how we design sample weights. We also show that the goal in Equation 1 can be derived from the machine learning optimization problem as described in this section.

By applying the sample weights for the local training on each client, the proposed method reduces the distribution skewness of each client’s data and prevents clients’ raw data from being exposed. Some statistical information between clients and the aggregator must be exchanged to find sample weights. However, instead of exchanging the raw information, which might hurt clients’ privacy, the proposed method only exchanges model parameters, similar to a typical Fl framework. ...

”””

Comment 1.4 Elaborate the effectiveness of the proposed method in reducing non-IID data impacts as compared to the existing methods in terms of accuracy, time and space complexity, and improvement in overall performance.

Response 1.4: Thank you for your question. To have a comprehensive comparison, in this revision, we have added a subsection, namely ”Discussion,” to discuss the proposed method’s performance over others in terms of accuracy, time, and space complexity. We also suggest the method to be used in cross-silo FL scenarios, which typically involve clients with high computing resources (in contrast to cross-device scenarios typically involving users with lightweight devices like mobile).

The main changes can be seen at page 3, line 7 and also listed as follows.

“‘ ....

A. Discusion

The proposed method offers a holistic improvement over existing federated learning methods. Its combination of enhanced accuracy and reduced communication costs signifies its effectiveness across a variety of datasets and scenarios. For example, the accuray can be increased by 22% and the communication time reduced by 8 times for FEMNIST dataset. While FedDisk exhibits remarkable performance across various metrics in federated learning scenarios, it’s important to acknowledge a drawback associated with its larger model size compared to some other methods as described in Table II. The increased model size leads to higher space occupation on client devices, which can have implications for light weight devices with limited storage capacity. Hence, the suggested approach could be well-suited

for the cross-silo scenario, typically characterized by clients having ample data and adequate computational capabilities. Overal, the proposed approach provides an avenue to address critical challenges in federated learning, making it a promising option for real-world applications. ...

”””

Comment 1.5 Provide complete details of the testbed set-up and the way datasets are used during experimentation within the manuscript to allow replication and verification of this work for the further growth of this research.

Response 1.5: Thank you for your suggestion. In this revision, additional to the detail on how to use the data in Experiments section, we also added an algorithms section which contains information on how to use the data in each step. Further more, we published our code and datasets to provide audiance a tool for reproducing the work. The github link is attached in the manuscript footnote. Link for code and data: ”https://github.com/nsh135/FedDiskPytorch”.

The algorithm can be seen at Section III-A and also listed as follows.

“‘ ....

E. Algorithm

Our main strategy can be described in Algorithm 3. We only concentrated on describing the first phase of FedDisk as the second is the same as a typical FL framework. First, each client trains its own local MADE model on the local data to obtain local distribution information (lines $1-3$). All clients then jointly train global MADE utilizing a typical FL framework (lines $4-14$). After achieving these two models, data are sampled from the two output models to acquire data samples containing local and global information (lines $15-17$). These samples are concatenated with the pseudo label of 0 for samples that come from local distribution and 1 for ones from the global distribution (line 18). They are then used for training an adversarial binary classifier (line 19). The purpose is to differentiate the two datasets sampled from the two distributions. The samples that are similar to the global samples will return a higher probability of belonging to class 1 (come from global distribution) and vice versa. Thus, the classifier probability-like output that represents class 1 is then used to be the weights for the sample (line 20).

...

”””

Comment 1.6 Please include following references within the manuscript at an appropriate place: 1 Wang, Zhibin, Jiahang Qiu, Yong Zhou, Yuanming Shi, Liqun Fu, Wei Chen, and Khaled B. Letaief. ”Federated learning via intelligent reflecting surface.” IEEE Transactions on Wireless Communications 21, no. 2

Algorithm 2 Phase 1: Sample Weight Computing.

Input: Client k^{th}: Dataset $\{X_k, y_k\}$.
Parameter: K : Number of client.
N : Number of total samples.
N_k: Number of sample of client k^{th}. $\sum_{k=1}^{K} N_k = N$
$local_iter$: Number of iterations for training local model
$global_iter$: Number of iterations for training global model.
$ld(\hat{w})$: Local MADE model containing **l**ocal **d**istribution information
$gd(\tilde{w})$: Global MADE model containing **g**lobal **d**istribution information
$h(\bar{w}_k)$: Shallow Binary Classifier to differentiate output from p_x and q_x α_k: sample weights for client k data
Output: α_k: Sample weights for X_k

{Training local MADE model}
for $k \leftarrow 1$ to K **do**
- Client k:

  Fully train $ld_k(\hat{w_k})$ on local data $\{X_k, y_k\}$.
end for

{Training global MADE model}
for $i \leftarrow 1$ to $global\_iter$ **do**
  for $k \leftarrow 1$ to K **do**
  - Client k:

    Update $\tilde{w}^{i-1}$ from the Aggregator
    for $j \leftarrow 1$ to $local\_iter$ **do**
      Train $gd(\tilde{w})$ on $\{X_k, y_k\}$.
    end for
    Sending $gd(\tilde{w})$ to Agreegator
  end for
  - Aggregator:

    Aggregate $\tilde{w}^i = \sum_{k=1}^{K} \frac{N_k}{N} \tilde{w}_k^i$
    Broatcasting $\tilde{w}^i$ to clients
end for

{Training shallow binary classifier}
for $k \leftarrow 1$ to K **do**
  Sample data from local distribution:
  $X_k^{\prime local} \leftarrow ld(X_k|\hat{w}_k)$, $y_k^{\prime local} \leftarrow [0...0]$
  Sample data from global distribution:
  $X_k^{\prime glob} \leftarrow gd(X_k|\tilde{w}_k)$, $y_k^{\prime global} \leftarrow [1...1]$
  $X_k' \leftarrow concat(X_k^{\prime local}, X_k^{\prime glob})$, $y_k' \leftarrow concat(y_k^{\prime local}, y_k^{\prime global})$
  - Client k:

  Fully train $h(\bar{w}_k)$ on local data $\{X_k', y_k'\}$.

{Estimate sample weight}
  $\alpha_k \leftarrow h(X_k'|\bar{w})[:1]$
end for

(2021): 808-822. Singh, Ashutosh Kumar, Deepika Saxena, Jitendra Kumar, and Vrinda Gupta. "A quantum approach towards the adaptive prediction of cloud workloads." IEEE Transactions on Parallel and Distributed Systems 32, no. 12 (2021): 2893-2905. Lim, Wei Yang Bryan, Jer Shyuan Ng, Zehui Xiong, Jiangming Jin, Yang Zhang, Dusit Niyato, Cyril Leung, and Chunyan Miao. "Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning." IEEE Transactions on Parallel and Distributed Systems 33, no. 3 (2021): 536-550.

Response 1.6: Thank you for your suggestion. In this revision, we have considered all the suggested works and mentioned in the Related Work section.

The work references can be found at [39], [9] and [10] respectively.

Comment 1.7 Please carefully proof-read the complete manuscript for typos and English grammar before the revised manuscript submission.

Response 1.7: Thank you for your suggestion. In this revision, we have corrected a number of typos and grammar errors.

III. RESPONSE TO REVIEWER 2

Comment 2.1 The introduction is too long and complicated, which includes the part of related work actually. You should separate the Introduction and Related Work into two different sections.

Response 2.1: Thank you for your suggestion. In this revision, we have revised the introduction and devide the section into ”Introduction” and ”Related Work” sections. The two sections have been rewritten to improve the comprehensive of the manuscript.

Comment 2.2 In the introduction, please emphasize your contribution and innovation by using simple language, not to list a lot of references.

Response 2.2: Thank you for your suggestion. In this revision, we have revised the introduction section and split the related work into a separate section. The introduction breifly introduce the problem, our distribution and the inovation.

Our contribution and inovation is discussed in the introduction section as follows:

“‘ ... To address this problem, we proposed an algorithm that utilizes sample weights to adjust individual client distributions closer to the global distribution during the training process. However, obtaining global information across clients is challenging in an FL setting because clients need to allow the exposure of their raw data. To overcome this challenge, the proposed method implicitly shares statistical information of client data without revealing the client’s raw data. The method only requires clients to exchange additional model weights using a typical FL procedure. Once the adjustment weights are acquired, the machine learning model can be trained using a standard FL framework. The proposed method is demonstrated to improve FL accuracy and significantly reduce FL communication costs through experiments on three real-world datasets.

Our contributions are as follows:

1) Provide a theoretical base for skewed feature distribution data for federated learning by adjusting sample weights derived from the machine learning empirical risk.
2) Provide a practical solution to mitigate the problem of learning from non-IID data for the FL framework without sharing clients’ draw data. The proposed method only requires clients to share additional model parameters, similar to a typical federated learning framework.
3) Several experiments were conducted on three datasets, including MNIST, non-IID benchmark dataset FEMNIST and real-world dataset Chest-Xray. The results demonstrate that the proposed method outperforms other experimental methods in classification accuracy and dramatically reduces the communication cost.

4) As the proposed method needs to exchange additional information, we also provide a theoretical analysis to analyze the potential privacy leakage. We showed that the leakage information becomes insignificant when the number of clients increases.
5) To our best knowledge, the proposed method is the first method utilizing data distribution information and sample weights to tackle the FL Non-IID issue.

... ”””

Comment 2.3 Please give enough recent references related to heterogeneous FL in the related work, such as 1 Wang, Mingjie, Jianxiong Guo, and Weijia Jia. ”FedCL: Federated Multi-Phase Curriculum Learning to Synchronously Correlate User Heterogeneity.” arXiv preprint arXiv:2211.07248 2022. 2 Zhu, Zhuangdi, Junyuan Hong, and Jiayu Zhou. ”Data-free knowledge distillation for heterogeneous federated learning.” International conference on machine learning. PMLR, 2021. 3 Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., & Feng, J. 2021. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. Advances in Neural Information Processing Systems, 34, 5972-5984. Especially for 3, it is very similar to your proposed method.

Response 2.3: Thank you for your suggestions. In this revision, we have considered all the above suggested works in the Related Work section. The coresponding references are [29], [20] and [19].

Comment 2.4 The problem definition and method are not clear. It is better to combine Section II and Section III, and give a formalized

Response 2.4: Thank you for your suggestion. In this revision, we have improved the comprehense of the manuscript by rewriting the Scenario section and change the section title to ”Problem Statment”.

The main changes can be seen at page 3, line 3 and also listed as follows.

“‘ ...

III. Problem Statement

In this section, we introduce and formulate the scenario of FL with skewed feature distribution across clients. Our scenario is a learning collaboration between K clients to build a global classification model that maximizes the global accuracy given arbitrary data. Each client holds a number of individual records that they are not willing to share with others due to privacy concerns. This study focuses on preventing the performance of the global model from deteriorating because of the distribution skewness issue [31] across clients.

Our goal is to adjust the clients' distributions to be closer to the global distribution via sample weights. We denote the data and associated labels held by client $k \in \{1, ..., K\}$ as $\{(\mathbf{x}_k^j, y_k^j)\}_{j=1}^{N_k}$ where $\mathbf{x}_k^j \in \mathbb{R}^d$

and $y_k^j \in \mathbb{N}$. Let the data distribution of the k^{th} client be $q_k(\mathbf{x})$ and the ground truth global distribution is $p(\mathbf{x})$. Our problem becomes finding an adjusting weight function for each client k, $\alpha(\mathbf{x}_k)$ such that

$$\alpha_k(\mathbf{x})q_k(\mathbf{x}) = p(\mathbf{x}) \tag{1}$$

”””

Comment 2.5 It is better to give a detailed description of your training method, such as ”algorithm” structure. This is imperative to thoroughly understand your idea.

Response 2.5: Thank you for your suggestion. In this updated version, we've included a new subsection titled ”Algorithm” to enhance the comprehensiveness of the manuscript.

The algorithm section can be seen at Section III-A and also listed as follows.

“‘ ....

A. Algorithm

Our main strategy can be described in Algorithm 3. We only concentrated on describing the first phase of FedDisk as the second is the same as a typical FL framework. First, each client trains its own local MADE model on the local data to obtain local distribution information (lines $1-3$). All clients then jointly train global MADE utilizing a typical FL framework (lines $4-14$). After achieving these two models, data are sampled from the two output models to acquire data samples containing local and global information (lines $15-17$). These samples are concatenated with the pseudo label of 0 for samples that come from local distribution and 1 for ones from the global distribution (line 18). They are then used for training an adversarial binary classifier (line 19). The purpose is to differentiate the two datasets sampled from the two distributions. The samples that are similar to the global samples will return a higher probability of belonging to class 1 (come from global distribution) and vice versa. Thus, the classifier probability-like output that represents class 1 is then used to be the weights for the sample (line 20).

...

”””

Comment 2.6 In the experiment part, this is not enough ablation experiment to verify your proposed module.In addition, it is better to compare with the more recent development in this area.

Response 2.6: Thank you for your recommendation. In this revised iteration, we've integrated a fresh subsection labeled ”Ablation Study: FedDiskAb.” Within this section, we explore a scenario where the aggregator could potentially hold the complete dataset, allowing for the direct derivation of sample weights

Algorithm 3 Phase 1: Sample Weight Computing.

Input: Client k^{th}: Dataset $\{X_k, y_k\}$.
Parameter: K : Number of client.
N : Number of total samples.
N_k: Number of sample of client k^{th}. $\sum_{k=1}^{K} N_k = N$
$local_iter$: Number of iterations for training local model
$global_iter$: Number of iterations for training global model.
$ld(\hat{w})$: Local MADE model containing **l**ocal **d**istribution information
$gd(\tilde{w})$: Global MADE model containing **g**lobal **d**istribution information
$h(\bar{w}_k)$: Shallow Binary Classifier to differentiate output from p_x and q_x α_k: sample weights for client k data
Output: α_k: Sample weights for X_k

{Training local MADE model}
for $k \leftarrow 1$ to K **do**
- Client k:

  Fully train $ld_k(\hat{w_k})$ on local data $\{X_k, y_k\}$.
end for

{Training global MADE model}
for $i \leftarrow 1$ to $global\_iter$ **do**
  for $k \leftarrow 1$ to K **do**
  - Client k:

    Update $\tilde{w}^{i-1}$ from the Aggregator
    for $j \leftarrow 1$ to $local\_iter$ **do**
      Train $gd(\tilde{w})$ on $\{X_k, y_k\}$.
    end for
    Sending $gd(\tilde{w})$ to Agreegator
  end for
  - Aggregator:

    Aggregate $\tilde{w}^i = \sum_{k=1}^{K} \frac{N_k}{N} \tilde{w}_k^i$
    Broatcasting $\tilde{w}^i$ to clients
end for

{Training shallow binary classifier}
for $k \leftarrow 1$ to K **do**
  Sample data from local distribution:
  $X_k^{\prime local} \leftarrow ld(X_k|\hat{w}_k)$, $y_k^{\prime local} \leftarrow [0...0]$
  Sample data from global distribution:
  $X_k^{\prime glob} \leftarrow gd(X_k|\tilde{w}_k)$, $y_k^{\prime global} \leftarrow [1...1]$
  $X_k' \leftarrow concat(X_k^{\prime local}, X_k^{\prime glob})$, $y_k' \leftarrow concat(y_k^{\prime local}, y_k^{\prime global})$
  - Client k:

  Fully train $h(\bar{w}_k)$ on local data $\{X_k', y_k'\}$.

{Estimate sample weight}
  $\alpha_k \leftarrow h(X_k'|\bar{w})[:1]$
end for

from the data. Through the replacement of this weight computation aspect, we've conducted tests to verify the effective functioning of our distribution learning via the MADE model. The results of the experiment involving FedDiskAb underscore its superior performance compared to alternatives, thus reinforcing the concept of utilizing sample weights to effectively address the challenge posed by non-IID data.

In addition, this revision we have added another state-of-the-art method, namely FedPCL, for a more comprehensive comparision. All the experiements have been updated with the new compared method.

The ”Ablation Study: FedDiskAb” section can be viewed at Section IV and also listed as follows.

“‘ ....

IV. Ablation Study: FedDiskAb

In this section, we examine the idea of using sample weight for non-IID data and the FedDisk sample weight effectiveness by looking at the case when the weights are derived directly from the raw data. Specifically, instead of learning sample weights from the local and global MADE model output, the weights are learned directly from the raw local and global data. To obtain the global data, we combine all client's data and randomly sample the same number of the client dataset size to avoid data imbalance. This setting variant of FedDisk (namely FedDiskAb) is an ideal case for a sample weight-based approach as it assumes to have access to the raw data. To obtain sample weights, FedDiskAb only needs to train the binary classifier on the combination of local data and global data, aiming to discriminate the two datasets. The classifier's output is used to derive the sample weight, similar to FedDisk.

Several experiments have been conducted for FedDiskAb and other methods in Section VII. The outcome demonstrates that both FedDiskAb and FedDisk surpass the performance of all alternative methods. This confirms the effectiveness of the sample weight-based strategy, whether acquired through learning the distribution from MADE models or directly from the data, in enhancing federated learning when faced with non-IID challenges. Furthermore, the performance of FedDisk closely aligns with that of FedDiskAb. This shows the fact that the local and global MADE models contribute significantly to the framework's ability to capture essential distribution information, much akin to the process of direct learning from the raw data. The details of the experimental results will be shown in Chapter VII. ...

”””

Comment 2.7 For the reader to reproduce your work, please open your datasets and source code. In this area, if you cannot prove reproducibility, this is meaningless. Response 2.7: Thank you for your recommendation. We published our code and datasets to provide audiance a tool for reproducing the work. The github link is attached in the manuscript footnote. Link for code and data: ”https://github.com/nsh135/FedDiskPytorch”.