A DYNAMIC REWEIGHTING STRATEGY FOR FAIR FEDERATED LEARNING

Zhiyuan Zhao *Gauri Joshi*

Carnegie Mellon University, Pittsburgh, USA
{zhiyuan2, gaurij}@andrew.cmu.edu

ABSTRACT

Federated learning is an emerging machine learning framework where models are trained using heterogeneous datasets collected by a large number of edge clients. Standard methods to aggregate local training models weigh each model by a fraction of data size at that client. However, such approaches result in unfairness to clients with small and unique datasets, leading to inferior accuracy of the global model at these clients. In this work, we propose a novel optimization framework called DRFL that dynamically adjusts the weight assigned to each client, and we combine it with a biased client selection strategy, both of which encourage fairness in federated training. We validate the effectiveness of our proposed method on a suite of both synthetic and real federated datasets, revealing the proposed method outperforms existing baselines in terms of resulting fairness.

Index Terms— federated learning, fairness, distributed optimization

1. INTRODUCTION

Federated learning has emerged as an attractive paradigm for machine learning optimization problems, in which the training process executes on an extensive distributed system with massive clients or edge devices, and a central server aggregates local models to a global model that is leveraged by all existing and upcoming clients. Practically, clients may contain local data from different distributions, resulting in data heterogeneity [1, 2]. While various algorithms had been proposed addressed with federated optimization settings since FedAvg [3], handling data heterogeneity remains an open problem in federated learning [1, 4]. Due to the existence of data heterogeneity, a model trained on the local dataset at one client may not work well for another client. In federated learning, minorities or marginalized groups are typically under-represented in training data, and thus the global model tends to weigh these groups less during training [5, 6, 7].

In this work, we aim to mitigate the *representation disparity* [1] to improve the fairness of the global model obtained in federated learning among all groups, including the minorities. To address this, we propose a new optimization objective DRFL that dynamically adjusts the weight assigned to each client. We also combine a biased client selection strategy to further improve the resulting fairness. We introduce a new algorithm that resolves the optimization problem above, and our experimental results illustrate the proposed method outperforms FedAvg and q-FFL baselines in resulting fairness. In addition, another key advantage of our approach over previously proposed methods is that it does not change the local optimizer at each client, and only makes adjustments to the aggregation weights on the server-side.

2. PROBLEM FORMULATION

System Model. In the standard federated learning setup, there are total K clients, where for each client k, it contains a local dataset D_k with size s_k. The clients communicate model updates to a central aggregating server, which aims to find a model parameter vector $\boldsymbol{x}$ that minimizes the empirical risk objective:

$$F(\boldsymbol{x}) = \sum_{k=1}^{K} p_k F_k(\boldsymbol{x}) = \sum_{k=1}^{K} \frac{s_k}{\sum_{i=1}^{K} s_i} F_k(\boldsymbol{x}) \tag{1}$$

The standard algorithm to solve (1) is *federated averaging* (FedAvg) [3]. FedAvg is executed in communication rounds, where in each round, the central server selects only a fraction C of $m = \max\{CK, 1\}$ clients with probability distribution $p_k = \frac{s_k}{\sum_{i=1}^{K} s_i}$ for training. The selected clients perform $\tau_k = E\frac{s_k}{B}$ local updates, where B is the batchsize and E is the number of local epochs. The central server takes a weighted average of the clients' model updates (p_k for client k) and updates the global model.

Though FedAvg is communication-efficient and achieves high overall accuracy in experiments, the selection of clients and server aggregation with unbiased weights p_k can cause fairness issues. Specifically, clients with a larger dataset size s_k are optimized more often and better than those of clients with small s_k, which typically represent minorities or outliers, since they are less likely to be selected. To clarify what $fairness$ refers to in federated learning, we formally define the desired fairness criteria as:

Definition 2.1 (*Fairness in Federated Learning* [7])**.** For global model $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$, we say x is a more *fair* solution for

This project was supported by NSF CCF-2045694 and CCF-2107024.

defined federated learning objective than $\bar{\boldsymbol{x}}$ if the accuracy distribution of model $\boldsymbol{x}$ on m devices is more *uniform* than $\bar{\boldsymbol{x}}$.

Related Prior Work. Fairness is a topic with various open challenges that have received broad attention and contributions in the machine learning community [8, 9, 10, 11]. Regarding fairness in federated learning, multiple works [6, 7, 12] explores possible methods for better fairness in two general approaches: optimizing more often to clients with high losses or giving more penalties to clients with high losses. *Agnostic federated learning* (AFL) [6] is a method that leverages the first approach mentioned above, where it tries to optimize the worst-performing client by giving it the largest weight. AFL aims to minimize a modified objective function described as:

$$F_{\texttt{AFL}}(\boldsymbol{x}) = \max_{\lambda} \sum_{k=1}^{K} \lambda_k F_k(\boldsymbol{x}) \tag{2}$$

where $\lambda = (\lambda_1, \ldots, \lambda_K)$ lies in the simplex $\wedge = \{\lambda : \lambda_k \geq 0, \sum_{k=1}^{K} \lambda_k = 1\}$. Intuitively, AFL uses a biased client selection strategy that optimizes the model for the worst performing client in each round. Since it chooses a single client for optimization in each round, the gradient variance can be much larger than that of the weighted averaging from multiple selected clients, causing potential convergence stability issue.

The q-FFL framework [7] utilizes the second approach, in which the global objective is defined in a power form:

$$F_q(\boldsymbol{\omega}) = \sum_{k=1}^{K} \frac{p_k}{q+1} F_k^{q+1}(\boldsymbol{\omega}) \tag{3}$$

where $q \geq 0$ is a tunable parameter. By adding power terms to the objective function, local clients with high losses can be magnified, resulting in higher penalties despite their s_k can be small. [7] also proposed a FedAvg-liked algorithm q-FedAvg to solve its proposed objective, and leverages the estimated Lipschitz constant to avoid tuning learning rate for different q. However, choosing a proper q for various datasets can be extremely sensitive, for instance, it uses $q = 5$ for Vehicle dataset but $q = 0.001$ for Shakespeare dataset, which can cause challenges in practical implementation.

3. FAIR FEDERATED LEARNING

To address fairness challenges in federated learning, and to overcome some of the shortcomings of existing methods, we proposed a novel global objective DRFL that can dynamically adjust the weight assigned to clients. Additionally, we utilize an optimized biased client selection strategy than AFL, which promises better convergence stability. Finally, we derive a complete algorithm DR-FedAvg to solve DRFL and propose potential optimizations towards our algorithm.

Dynamic Reweighting Federated Learning. Building on the q-FFL method reviewed in Section 2, rather than assigning more penalties to clients with high losses through loss magnification, can we add penalties to these clients simply by assigning higher weights? Guided by this idea, we propose a global objective function, namely dynamic-reweighting federated learning objective (DRFL):

$$F_{\texttt{DRFL}}(\boldsymbol{x}) = \sum_{k=1}^{K} \frac{p_k F_k^{q+1}(\tilde{\boldsymbol{x}})}{\sum_{i=1}^{K} p_i F_i^{q+1}(\tilde{\boldsymbol{x}})} F_k(\boldsymbol{x}) \tag{4}$$

where $q \geq 0$ is a tunable parameter similar to q-FFL. $\boldsymbol{x}$ is the model vector obtained from the current communication round and $\tilde{\boldsymbol{x}}$ is a stale model from previous training, i.e. the model obtained from the last communication round. To avoid tricky parameter tuning, a common strategy is to set $q = 0$, each client k is assigned with a weight based on its performance in the previous step, that is, its loss on a stale model normalized by the sum loss of all clients. Suppose the loss for each client with a global model obtained in the previous round can be calculated or estimated at the beginning of every training round, then we could dynamically adjust penalties for each client. Clients with higher loss are assigned with large weights, receiving more penalties, and clients with lower loss work contrarily. If a more fair model is desired, one can set $q > 1$ to give even heavier penalties on bad clients. Also, by letting $q = -1$, (4) reduces to vanilla FedAvg.

Comparing to q-FFL, DRFL utilizes similar mechanism as q-FFL, which exaggerates the impacts on clients with high losses. However, DRFL is more flexible in parameter tuning, where one can simply set $q = 0$ for simplicity, or tuning q for the desired fairness. The proposed q-FedAvg to solve q-FFL also requires computing Δ_k and h_k in the client-side besides the gradient calculations, causing potential aggregation information leaking and harming system security. Meanwhile, q-FFL requires to estimate a Lipschitz constant at $q = 0$ for a replacement of learning rate, which causes large extra computation cost during training.

Biased Client Selection. Since DRFL requires a loss estimation over clients for their dynamic weights, one can take convenience from loss estimation to execute a biased client selection. To overcome unstable convergence occurred in AFL, Pow-D [13] proposed an optimized client selection strategy, in which it firstly selects d clients to obtain their loss $\{F_k(x), k \in d\}$, then it picks top $m = \max\{CK, 1\} \leq d$ clients with highest losses to perform local updates. Since Pow-D counts an averaged gradients, it is rational that Pow-d performs better than AFL in stability.

Our clients selection strategy is similar to Pow-d, in which we estimate losses over selected d clients based on p_k, and pick picks top $m = \max\{CK, 1\}$ clients with highest losses to perform local updates. However, when the total number of clients is relatively small, we tend to choose $d = K$ for better generality since it evaluates the performance of all clients rather than a partition. When the client amount is massive and letting $d = K$ is not practical, we back to pick a fraction of clients as Pow-D does.

Dynamic Reweighting Federated Averaging. To resolve the proposed objective function, we propose a FedAvg-style algorithm names *Dynamic-Reweighting Federated Averaging* (DR-FedAvg). The DR-FedAvg takes advantages of communication efficiency from FedAvg, combines the above client selection strategy to optimize DRFL objective function, guides the global model to become more fair. The full algorithm is described in Algorithm 1.

Algorithm 1 Dynamic-Reweighting Federated Averaging

```
Initialization: $T, K, d, C, \tau, q$
for $t$ in $\{0, \dots, T-1\}$ rounds do
    Server samples $d$ clients with probability $p_k$, and sends current model $\tilde{\boldsymbol{x}}$ to clients to obtain estimated loss $\hat{F}_k(\tilde{\boldsymbol{x}})$ over a batch local data;
    Clients send $\hat{F}_k(\tilde{\boldsymbol{x}})$ to server, and server picks a subset $s_t$ of $m = \max\{CK, 1\}$ clients with highest losses;
    Selected clients $k \in s_t$ perform $\tau$ local updates to obtain $\boldsymbol{x}$ and send back to server;
    Server calculates dynamic weights for updated clients $w_k = \frac{p_k F_k^{q+1}(\tilde{\boldsymbol{x}})}{\sum_{i=1}^{m} p_k F_k^{q+1}(\tilde{\boldsymbol{x}})}$, $k \in s_t$ for current round, and aggregates models by $\boldsymbol{x} \leftarrow \sum_{k=1}^{m} w_k \boldsymbol{x}_k$
end for
return $\boldsymbol{x}$
```

To save computation resources, a potential optimization for DR-FedAvg could be: rather than updating the dynamic weights for each round, one can update the weights for every t rounds. To adapt the biased client selection, when the weights are to update, one might pick a relatively larger d (client set S_d), obtain their estimated loss, and perform Pow-D-like updates with biased client selection. For the incoming $t - 1$ rounds, one can sample $m = \max\{CK, 1\}$ clients only from S_d to perform vanilla FedAvg. By so, the dynamic weights only need to be updated for every t rounds.

4. EXPERIMENT EVALUATION

We evaluate the effectiveness of proposed method. For comparison, we also implement FedAvg and q-FedAvg as baselines. The experiments are divided into two parts: First, we evaluate DR-FedAvg versus baseline methods over synthetic datasets, illustrating the proposed method leads to a more uniform distribution; Second, we evaluate all models over Adult and FMNIST dataset to show that DR-FedAvg improves the performance over minorities with slightly sacrificing global accuracy or performance over majorities.

Evaluation over Synthetic Data. For synthetic data, we use SYNTHETIC(α, β) that follows same settings as [14]. α, β are two varying parameters, in which α controls how much local models are different from each other, and β controls how much local data at each device differs from that of other devices. Intuitively, a larger set of α, β indicates a less IID dataset it generates. In our experiment, we leverage SYNTHETIC$(0, 0)$ and SYNTHETIC$(1, 1)$ with each dataset contains 100 clients, and build single-layer perceptrons $W_{(0,0)}^{60\times5}$ and $W_{(1,1)}^{60\times2}$ respectively. The goal of this set of experiments is not to reach as high accuracy as possible, but to compare different performance distributions over different methods under a similar global accuracy, thus, whether a model is large enough does not matter too much. Since the clients group is relatively small, we set $d = K$ and $q = 0$ for DRFL. For consistency and simplicity, we set local update epoch $\tau = 1$ and *lr* $= 0.01$ using SGD optimizer for all methods. The experiment results are shown in Fig 4.

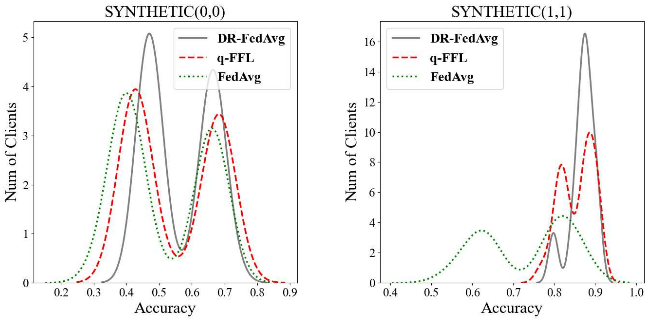

Fig. 1. Performance distribution on SYNTHETIC(0,0) (left) and SYNTHETIC(1,1) (right), showing a more uniform performance on DR-FedAvg than baselines

Recall the Definition 2.1, the fairer the model is, the more uniform the accuracy distribution should be, and the accuracy distribution should be more concentrated in the plotting figure. For SYNTHETIC$(0, 0)$, all methods saturate at approximate 60% global accuracy. However, either FedAvg or q-FFL has two splits over accuracy range 30% $\sim$ 50% and 60% $\sim$ 70%, while performances distribution with DR-FedAvg is slight more concentrated, indicating a better fairness than baselines. For *SYNTHETIC*$(1, 1)$, the task is simplified to a binary classification problem such that all methods achieves higher global accuracy than *SYNTHETIC*$(0, 0)$ (not equivalent to *SYNTHETIC*$(0, 0)$ is less iid than *SYNTHETIC*$(1, 1)$). The resulting fairness of DR-FedAvg is slightly better than q-FFL. Experiment also shows that FedAvg saturates at 80% global accuracy while other two methods reach higher accuracy, which is due to a poor performance of FedAvg over non-IID dataset, and has been illustrated by previous works [15]. The experiments over synthetic dataset demonstrate two points:1) DR-FedAvg can work well over non-IID data, 2) DR-FedAvg has a better guarantee of worst-case than both baselines.

Evaluation over Real Data. For real dataset, we leverage

Dataset	Adult		
Method	Avg	PhD	non-PhD
FedAvg	83.1	65.6± 3.9	83.3± 0.1
q-FFL	83.3± 0.1	75.7± 4.7	83.4± 0.1
DR-FedAvg	83.1± 0.1	75.9± 3.1	83.2± 0.1

Dataset	FMNIST			
Method	Avg	Shirt	Pullover	T-shirt
FedAvg	77.7± 3.2	**70.7**± 13.0	77.4± 9.5	85.3± 8.3
q-FFL	82.9± 1.6	**78.5**± 11.5	85.0± 9.4	83.2± 8.4
DR-FedAvg	81.4± 1.2	80.5± 3.6	82.1± 4.1	**79.1**± 3.6

Table 1. Validation Accuracy (%) over Adult and FMNIST, DR-FedAvg achieves highest accuracy over PhD (minority) on Adult, and most uniform accuracy distribution on FMNIST

Adult [16] and FMNIST [17]. Following a similar setup in [7], for Adult, we split the dataset into PhD and non-PhD groups, where PhD group is the minority that receives much lower prediction accuracy compared to the global average. For FMNIST, we sample data from categories Shirt, Pullover, and T-shirt. Our goal is to evaluate whether DR-FedAvg could improve the performance of PhD of Adult, and Shirt of FMNIST than FedAvg as q-FFL does, with rational scarification of accuracy over other groups.

In the experiments, we build a 3-layer DNN for the Adult, and 2-layer convolutional layers with a 2-layer MLP classifier followed for FMNIST. We do non-IID sampling for both datasets, for each client, it contains data and labels only from a single category. We employ 30 clients for the Adult, with one client containing all data from PhD group, and another 29 with data sampled from non-PhD group without replacement. For FMNIST, we leverage 60 clients, with 20 for each selected category with relatively uniform samplings. All other settings are the same as experiments over the synthetic dataset, the results are shown in Fig 4 and Table 4.

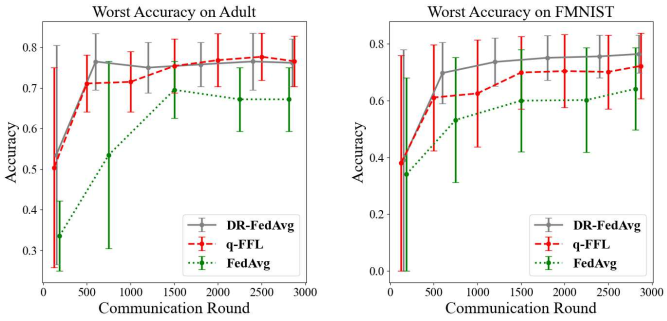

Fig. 2. Worst category performance on Adult (left) and FMNIST (right) over training rounds, DR-FedAvg ensures better guarantee over all categories than baselines

Dataset	Adult		
q-value	Avg	PhD	non-PhD
0	83.0± 0.1	74.6± 2.5	83.1± 0.1
0.1	83.1± 0.1	75.1± 2.5	83.1± 0.1
1	83.3± 0.1	**75.6**± 2.2	83.4± 0.1
10	82.9± 0.1	71.7± 3.1	83.0± 0.1

Dataset	FMNIST			
q-value	Avg	Shirt	Pullover	T-shirt
0	85.4± 0.9	86.4± 6.6	90.1± 4.4	79.7± 6.1
0.1	84.7± 0.8	86.2± 6.6	87.7± 4.9	80.2± 7.1
1	84.9± 0.8	85.5± 6.7	88.6± 5.6	**80.7**± 7.6
10	84.6± 0.5	85.2± 5.8	88.9± 4.4	79.8± 7.5

Table 2. Validation Accuracy (%) over Adult and FMNIST of DR-FedAvg with different q values, indicating $q = 1$ gives best worst-case guarantees on both Adult and FMNIST.

Fig 4 reveals that DR-FedAvg achieves a faster convergence speed than the baselines, and DR-FedAvg or q-FFL have better guarantees on minorities than vanilla FedAvg. The proposed method performs better than baselines on FMNIST rather than Adult, which possibly due to a heterogeneity in client datasize in Adult sampling, where a relatively small p_k for PhD group degrades its contribution during aggregation. The figure also indicates a smaller confidence interval of DR-FedAvg than other methods. Table 4 further indicates that, DR-FedAvg ensures a even more fair result and a much better stability guarantee than q-FFL, though its global accuracy is slightly lower than q-FFL.

In addition, we also investigate the impact of q by switching different values. Intuitively, a larger q can ensure better fairness since it exaggerates more penalties than smaller q does. However, our empirical result, shown in Table 4, does not indicate a larger q is always favorable. The evaluation shows that DR-FedAvg has best worst-case guarantees when $q = 1$ on both Adult and FMNIST, and our interpretation is that a large q can cause higher error flow and instability, which may sacrifice the global accuracy too much and degrade the performance of all categories, including minorities.

5. CONCLUSION

In this work, we proposed DRFL, a novel optimization objective that encourages more uniform accuracy distributions across devices in federated learning. We devise a method DR-FedAvg that can solve the proposed objective efficiently in massive networks. Our empirical evaluation on a suite of the federated datasets, including synthetic and real datasets, demonstrates our method can achieve better fairness than baselines and can avoid sensitive parameter tunings in q-FFL. Future works may make extensions in two directions: 1) Convergence analysis for DRFL to support its feasibility theoretically, 2) Giving theoretical proofs of how DR-FedAVg leads to better fairness, i.e. with DR-FedAVg, the variance over clients is theoretically lower than FedAvg.

6. REFERENCES

[1] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[2] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor, "A novel framework for the analysis and design of heterogeneous federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5234–5249, 2021.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[4] 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," *arXiv preprint arXiv:1912.04977*, 2019.

[5] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[6] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[7] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[8] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[9] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3995–4004.

[10] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[11] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[12] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6357–6368.

[13] 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.

[14] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[15] 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.

[16] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[17] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017.