Federated Learning with Fair Averaging

Zheng Wang[1], **Xiaoliang Fan**[1,*], **Jianzhong Qi**[2],
Chenglu Wen[1], **Cheng Wang**[1] and **Rongshan Yu**[1]
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China
[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
zwang@stu.xmu.edu.cn,
fanxiaoliang@xmu.edu.cn, jianzhong.qi@unimelb.edu.au, {clwen, cwang, rsyu}@xmu.edu.cn

Abstract

Fairness has emerged as a critical problem in federated learning (FL). In this work, we identify a cause of unfairness in FL – *conflicting* gradients with large differences in the magnitudes. To address this issue, we propose the federated fair averaging (FedFV) algorithm to mitigate potential conflicts among clients before averaging their gradients. We first use the cosine similarity to detect gradient conflicts, and then iteratively eliminate such conflicts by modifying both the direction and the magnitude of the gradients. We further show the theoretical foundation of FedFV to mitigate the issue conflicting gradients and converge to Pareto stationary solutions. Extensive experiments on a suite of federated datasets confirm that FedFV compares favorably against state-of-the-art methods in terms of fairness, accuracy and efficiency. The source code is available at https://github.com/WwZzz/easyFL.

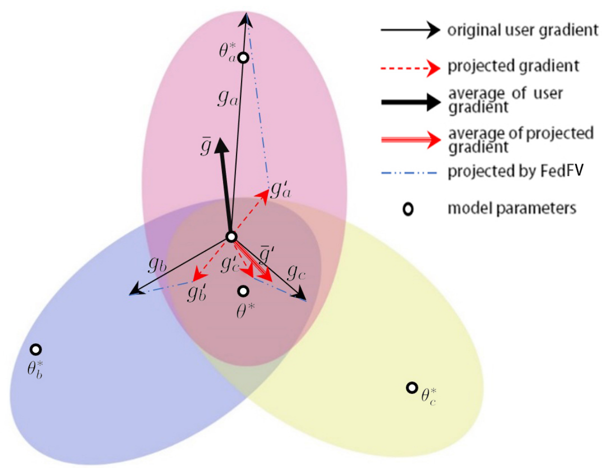


Figure 1: When there are conflicting gradients ($g_a \cdot g_b{<}0, g_a \cdot g_c{<}0, g_b \cdot g_c{<}0$) and large differences in the gradient magnitudes, $||g_a||{>}||g_b||{>}||g_c||$, the average of the original gradients $\bar{\theta}$ will be dominated by g_a, which will be far away from the global optimal θ^*, resulting in unfairness to the dominated clients b and c. We argue that if projecting the gradients to mitigate the conflicts before averaging, the update direction $\bar{g}'$ will be corrected, which is closer to the optimal and fairer.

1 Introduction

Federated learning (FL) has emerged as a new machine learning paradigm that aims to utilize clients' data to collaboratively train a global model while preserving data privacy [McMahan *et al.*, 2017]. The global model is expected to perform better than any locally trained model, since it has much more data available for the training. However, it is difficult for the global model to treat each client fairly [Kairouz *et al.*, 2019; Li *et al.*, 2019; Mohri *et al.*, 2019]. For example, a global model for face recognition may work well for younger users (clients) but may suffer when being used by more senior users, as the younger generation may use mobile devices more frequently and contribute more training data.

Techniques have been proposed to address the fairness issue in FL. AFL [Mohri *et al.*, 2019] aims at good-intent fairness, which is to protect the worst-case performance on any client. This technique only works on a small network of dozens of clients because it directly treats each client as a domain, which may suffer in generalizability. Li *et al.* [2019] seek to balance the overall performance and fairness using a fair resource allocation method. Hu *et al.* [2020] find a common degrade direction for all clients without sacrificing anyone's benefit and demonstrate robustness to inflating loss attacks. Li *et al.* [2020a] further explore a trade-off between a more general robustness and fairness, and they personalize each client's model differently.

Different from these studies, we observe that *conflicting* gradients with large differences in the magnitudes might fatally bring unfairness in FL. Since gradients are calculated locally by the clients, an update direction of some clients may hinder the model performance on other clients. The gradient conflicts seem not detrimental on their own since it has been shown that a simple average of gradients is valid to optimize the global objective function [Li *et al.*, 2020c]. However, when the conflicting gradients have large differences in their magnitudes, model accuracy can suffer significant reductions for some of the clients, as shown in Figure 1.

We highlight three characteristics of FL that make it critical in ensuring fairness for the clients.

*Corresponding Author

- **Heterogeneous and imbalanced data**. Different clients have their own data characteristics and hence potentially different data distributions. When the data is non-IID distributed, the divergence among the gradients of difference clients gets even larger [Zhao *et al.*, 2018]. These contribute conflicting gradients in FL. Further, different clients may also have different dataset sizes. If the same local batch size is used across the clients, a client with more data will take more steps to train the local model, thus leading to large differences in gradient magnitudes. As a result, the selected users in a communication round can largely conflict with each other, leading to an unfair update to some of them – we call such conflicts *internal conflicts*.

- **Party selection**. In each communication round, the server only randomly selects a subset of clients to train the global model with their local datasets. There is no guarantee that the data distribution in each round is representative of the real population distribution. The data distribution can be highly imbalanced when the server repeatedly chooses a particular type of clients, e.g., due to sampling strategies based on computing power, latency, battery etc., the average of gradients may become dominated by the repeatedly selected clients. Therefore, apart from the internal conflicts, we also identify *external conflicts* between those selected and those ignored by the server.

- **Client dropping out**. Another source of *external conflicts* comes from network interruption. When a client drops out due to network interruption (or other hardware issues), its magnitude is considered as 0. An update in the global model may also be unfair for the client.

To address the issues above, we propose **Fed**erated **F**air a**V**eraging (**FedFV**) algorithm to mitigate the conflicts among clients before averaging their gradients. We first use the cosine similarity to detect gradient conflicts and then iteratively eliminate such conflicts by modifying both the direction and magnitude of the gradients. We further show how well FedFV can mitigate the conflicts and how it converges to either a pareto stationary solution or the optimal on convex problems. Extensive experiments on a suite of federated datasets confirm that FedFV compares favorably against state-of-the-art methods in terms of fairness, accuracy and efficiency.

The contributions of this work are summarized as follows:

- We identify two-fold gradients conflicts (i.e., internal conflict and external conflict) that are major causes of unfairness in FL.

- We propose FedFV that complements existing fair FL systems and prove its ability to mitigate two gradient conflicts and converge to Pareto stationary solutions.

- We perform extensive experiments on a suite of federated datasets to validate the competitiveness of FedFV against state-of-the-art methods in terms of fairness, accuracy and efficiency.

2 Preliminaries

FL aims to find a shared parameter θ that minimizes the weighted average loss of all clients [McMahan *et al.*, 2017]:

$$\min_{\theta} F(\theta) = \sum_{k=1}^{K} p_k F_k(\theta) \tag{1}$$

where $F_k(\theta)$ denotes the local objective of the kth client with weight p_k, $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. The local objective is usually defined by an empirical risk on a local dataset, i.e., $F_k(\theta) = \frac{1}{n_k}\sum_{\xi_i \in D_k} l(\theta, \xi_i)$, where D_k denotes the local dataset of the kth client, and n_k is the size of D_k. To optimize this objective, McMahan *et al.* [2017] propose *FedAvg*, an iterative algorithm where the server randomly samples a subset S_t of m clients, $0<m \leq K$, to train the global model with their own datasets and aggregates local updates $G_t = \{g_1^t, g_2^t, ..., g_m^t\}$ into an average $g^t = \sum_{k=1}^{m} p_k g_k^t$ with weight $p_k = \frac{n_k}{\sum_{j=1}^{m} n_j}$ in the tth iteration. FedAvg has been shown to be effective in minimizing the objective with low communication costs while preserving privacy. However, it may lead to an unfair result where accuracy distribution is imbalanced among different clients [Li *et al.*, 2019; Mohri *et al.*, 2019].

In this paper, we aim to minimize Equation (1) while achieving a fair accuracy distribution among the clients.

3 Proposed Approach

We first present our approach, FedFV, for easing the negative impact of conflicting gradients with largely different magnitudes on FL fairness. We then give a theoretical analysis on how well FedFV achieves fairness and its convergence. We start by defining gradient conflicts following Yu et al. [2020].

Definition 1. *Client i's gradient $g_i \in R^d$ conflicts with client j's gradient $g_j \in R^d$ iff $g_i \cdot g_j<0$.*

3.1 FedFV

FedFV aims to mitigate the conflicts among clients before averaging their gradients. In the tth communication round of FedFV, after receiving the selected clients' updates $G_t = \{g_1^t, g_2^t, ..., g_m^t\}$ and training losses $L_t = \{l_1^t, l_2^t, ..., l_m^t\}$, the server updates the gradient history GH to keep a trace of the latest gradient of each client. Then, the server sorts gradients in G_t in ascending order of their corresponding clients' losses to obtain PO_t, which provides the order of each gradient to be used as a projection target. Finally, FedFV mitigates the internal conflicts and the external conflicts sequentially. Algorithm 1 summarizes the steps of FedFV.

3.2 Mitigating Internal Conflicts

We first handle the fairness among the selected clients in each communication round.

Definition 2. *In the tth communication round, there are internal conflicts if there is at least a pair of client gradients $< g_i^t, g_j^t >$ such that g_i^t conflicts with g_j^t, where $g_i^t, g_j^t \in G_t = \{g_1^t, g_2^t, ..., g_m^t\}$ and G^t is the selected clients' updates.*

Algorithm 1 FedFV

Input:$T, m, \alpha, \tau, \eta, \theta^0, p_k, k = 1, ..., K,$

Initialize θ_0 and gradient history $GH = []$.
for $t = 0, 1, ..., T-1$ **do**
  Server samples a subset S_t of m clients with the prob. p_k and sends the model θ^t to them.
  Server receives the updates g_k^t and the training loss l_k^t from each client $k \in S_t$, where $g_k^t \leftarrow \theta^t - \theta_k^t$ and θ_k^t is updated by client k. Then server updates the trace $GH = [g_1^{t_1}, g_2^{t_2}, ..., g_K^{t_K}]$ of the latest updates of all clients.
  Server sorts the clients' updates into a projecting order list $PO_t = [g_{i_1}^t, ..., g_{i_m}^t]$ based on their losses, where $l_{i_j}^t \leq l_{i_{j+1}}^t$.
  $g^t \leftarrow$MitigateInternalConflict(PO_t, G_t, α)
  if $t \geq \tau$ **then**
    $g^t \leftarrow$MitigateExternalConflict(g^t, GH, τ)
  end if
  $g^t = g^t/||g^t|| * ||\frac{1}{m}\sum_k^m g_k^t||$
  Server updates the model $\theta^{t+1} \leftarrow \theta^t - g^t$
end for

Algorithm 2 MitigateInternalConflict

Input: PO_t, G_t, α

Server selects a set S_t^α of the top α clients in PO_t.
Server sets $g_k^{PC} \leftarrow g_k^t$.
for each client $k \in S_t^\alpha$ in parallel **do**
  for each $g_{i_j}^t \in PO_t, j = 1, ..., m$ **do**
    if $g_k^{PC} \cdot g_{i_j}^t < 0$ and $k \neq i_j$ **then**
      Server computes $g_k^{PC} \leftarrow g_k^{PC} - \frac{(g_{i_j}^t) \cdot g_k^{PC}}{||g_{i_j}^t||^2} g_{i_j}^t$.
    end if
  end for
end for
$g^t \leftarrow \frac{1}{m}\sum_{k=1}^m g_k^{PC}$
return g^t

Internal conflicts account for the unfairness among selected clients. Consider learning a binary classifier. if clients with data of one class are in the majority in a communication round and there exist conflicts between gradients of these two classes, the global model will be updated to favor clients of the majority class, sacrificing the accuracy on clients of the other class. Further, even when the proportion of clients of the two classes is balanced, the magnitudes of gradients may still largely differ due to different dataset sizes on different clients.

To address the internal conflicts, FedFV iteratively projects a client's gradient onto the normal plane of another client with a conflicting gradient. Instead of projecting in a random order, we design a loss-based order to decide which gradient should be the projecting target of other gradients based on Theorem 1. We show that the later a client gradient is used as the projection target, the fewer conflicts it will have with the final average gradient computed by FedFV. Algorithm 2 summarizes our steps to mitigate internal conflicts.

Theorem 1. *Suppose there is a set of gradients $G = \{g_1, g_2, ..., g_m\}$ where g_i always conflicts with $g_j^{(t_j)}$ before projecting $g_j^{(t_j)}$ to g_i's normal plane, and $g_i^{(t_i)}$ is obtained by projecting g_i to the normal planes of other gradients in G for t_i times. Assuming that $|\cos < g_i^{(t_i)}, g_j^{(t_j)} >| \leq \epsilon, 0<\epsilon \leq 1$, for each $g_i \in G$, as long as we iteratively project g_i onto g_k's normal plane (skipping g_i it self) in the ascending order of k where $k = 1, 2, ..., m$, the larger k is, the smaller the upper bound of conflicts between the average $\bar{g}' = \frac{1}{m}\sum_{i=1}^m g_i^{(m)}$ and g_k is.*

Proof. See Appendix A.1. □

Since projecting gradients favors for the gradient that serves as the projecting target later, we put the clients with larger training losses at the end of the projecting order list PO_t in the tth communication round to improve the model performance on the less well trained data. We allow αm clients to keep their original gradients to further enhance fairness. We will detail α in Section 3.4.

3.3 Mitigate External Conflicts

Due to party selection and client dropping out, there is a sample bias during each communication round in FL [Kairouz *et al.*, 2019]. A client that is not selected in the tth round can suffer a risk of being forgotten by the model whenever the combined update g^t conflicts with its imaginary gradient g_{img}^t. We can consider such clients to have a weight of zero in each round. However, we cannot directly detect conflicts for such clients because we have no access to their real gradients. To address this issue, we estimate their real gradients according to their recent gradients, and we call such gradient conflicts external conflicts:

Definition 3. *In the tth communication round, there are external conflicts if there is a client $h \notin S_t$ whose latest gradient g_h^{t-k} conflicts with the combined update g^t, where $0<k<\tau, 0<\tau<t$.*

An external conflict denotes the conflict between the assumed gradient of a client that has not been selected and the combined update g^t. We add extra steps to prevent the model from forgetting data of clients beyond current selection. We also iteratively project the update g^t onto the normal plane of the average of conflicting gradients for each previous round in a time-based order. The closer the round is, the later we make it the target to mitigate more conflicts between the update g^t and more recent conflicting gradients according to Theorem 1. We scale the length of the update to $||\frac{1}{m}\sum_k^m g_k^t||$ in the end because the length of all gradients can be enlarged by the projection. Algorithm 3 summarizes our steps to mitigate external conflicts.

3.4 Analysis

We analyze the capability of FedFV to mitigate gradient conflicts and its convergence. We show how FedFV can find a pareto stationary solution on convex problems.

Algorithm 3 MitigateExternalConflict

Input: g^t, GH, τ
for round $t-i, i=\tau, \tau-1, ..., 1$ **do**
  $g_{con} \leftarrow 0$
  for each client $k = 1, 2, ..., K$ **do**
    if $t_k = t - i$ **then**
      if $g^t \cdot g_k^{t_k} < 0$ **then**
        $g_{con} \leftarrow g_{con} + g_k^{t_k}$
      end if
    end if
  end for
  if $g^t \cdot g_{con} < 0$ **then**
    Server computes $g^t \leftarrow g^t - \frac{g^t \cdot g_{con}}{||g_{con}||^2} g_{con}$.
  end if
end for
return g^t

Theorem 2. *Suppose that there is a set of gradients $G = \{g_1, g_2, ..., g_m\}$ where g_i always conflicts with $g_j^{(t_j)}$ before projecting $g_j^{(t_j)}$ to g_i's normal plane and $g_i^{(t_i)}$ is obtained by projecting g_i to the normal planes of different gradients in G for t_i times. If $\epsilon_1 \leq |\cos < g_i^{(t_i)}, g_j^{(t_j)} >| \leq \epsilon_2, 0 < \epsilon_1 \leq \epsilon_2 \leq 1$, then as long as we iteratively project g_i onto g_k's normal plane (skipping g_i it self) in the ascending order of k where $k = 1, 2, ..., m$, the maximum value of $|g_k \cdot \bar{g}'|$ is bounded by $\frac{m-1}{m}(\max_i ||g_i||)^2 f(m, k, \epsilon_1, \epsilon_2)$, where $f(m, k, \epsilon_1, \epsilon_2) = \frac{\epsilon_2^2 (1-\epsilon_1^2)^{\frac{1}{2}} (1-(1-\epsilon_1^2)^{\frac{m-k}{2}})}{1-(1-\epsilon_1^2)^{\frac{1}{2}}}$.*

Proof. See Appendix A.2. □

According to Theorem 2, it is possible for FedFV to bound the maximum conflict for any gradient by choosing k to let $f(m, k, \epsilon_1, \epsilon_2) < \epsilon_2$ with any given $m, m \geq 2$, since the possible maximum value of conflicts between g_k and the original average gradient is $\max_k |g_k \cdot \bar{g}| \leq \frac{m-1}{m} \epsilon_2 (\max_k ||g_k||)^2$.

In practice, we mitigate all the conflicts that have been detected to limit the upper bound of gradient conflicts of clients. We allow αm clients with large training losses to keep their original gradients to further enhance fairness. When $\alpha = 1$, all clients keep their original gradients so that FedFV covers FedAvg, and when $\alpha = 0$, all clients are enforced to mitigate conflicts with others. Parameter α thus controls the degree of mitigating conflicts and can be applied to seek for a balance.

Next, we show that FedFV can find a pareto stationary point according Theorems 3 and 4.

Theorem 3. *Assume that there are only two types of users whose objective functions are $F_1(\theta)$ and $F_2(\theta)$, and each objective function is differentiable, L-smooth and convex. For the average objective $F(\theta) = \frac{1}{2} \sum F_i(\theta)$, FedFV with stepsize $\eta \leq \frac{1}{L}$ will converge to either 1) a pareto stationary point, 2) or the optimal θ^*.*

Proof. See Appendix A.3. □

Theorem 4. *Assume that there are m objective functions $F_1(\theta)$, $F_2(\theta)$,..., $F_m(\theta)$, and each objective function is differentiable, L-smooth and convex. If $\cos < \bar{g}, \bar{g}' > \geq \frac{1}{2}$ and $||\bar{g}|| \geq ||\bar{g}'||$ where $\bar{g}$ is the true gradient for the average objective $F(\theta) = \frac{1}{m} \sum F_i(\theta)$, then FedFV with step size $\eta \leq \frac{1}{L}$ will converge to either 1) a pareto stationary point, 2) or the optimal θ^*.*

Proof. See Appendix A.3. □

4 Related Work

4.1 Fairness in FL

Federated Learning (FL) is first proposed by McMahan et al [2017] to collaboratively train a global model distributedly while preserving data privacy[Kairouz *et al.*, 2019; Li *et al.*, 2020b; McMahan *et al.*, 2017]. A number of studies focus on collaborative fairness where the server allocates different models to clients according to their contribution [Lyu *et al.*, 2020; Lyu *et al.*, 2019; Xu and Lyu, 2020]. A few other studies address the fairness of a uniform accuracy distribution across devices [Cho *et al.*, 2020; Hu *et al.*, 2020; Li *et al.*, 2019; Li *et al.*, 2020a; Mohri *et al.*, 2019]. Mohri *et al.* [2019] and Li *et al.* [2019] propose different federated objectives AFL and q-FFL to further improve fairness. Hu *et al.* [2020] observe the competing relation between fairness and robustness to inflating loss attacks. Abay *et al.* [2020] analyze the potential causes of bias in FL which leads to unfairness, and they also point out the negative impact of sample bias due to the party selection. Cho *et al.* [2020] also show that client selection has an impact on fairness, and they propose a bandit-based communication-efficient client selection strategy to overcome biasness. Huang *et al.* [2020] reweight clients according to their accuracy and numbers of times of being selected to achieve fairness. They use double momentum to accelerate the convergence. Different from these, we identify the conflicts among clients to be a potential cause of unfairness in FL. We mitigate such conflicts by computing a fairer average of gradients to achieve fair model performance across devices.

4.2 Gradient Projection

Gradient projection has been well studied in continual learning to mitigate the adversary impact of gradient updates to previously learned tasks [Chaudhry *et al.*, 2018; Farajtabar *et al.*, 2019; Guo *et al.*, 2019; Lopez-Paz and Ranzato, 2017]. Lopez-Paz *et al.* [Lopez-Paz and Ranzato, 2017] project gradient by solving a quadratic programming problem. Chaudhry *et al.* [2018] project gradient onto the normal plane of the average gradient of previous tasks. Farajtabar *et al.* [2019] project the current task gradients onto the orthonormal set of previous task gradients. Yu *et al.* [2020] focus on adversary influence between task gradients when simultaneously learning multiple tasks. They iteratively project each task gradient onto the normal plane of conflicting gradients, which motivates our solution in this paper. To the best of our knowledge, we are the first to take the adversary gradient interference into consideration in FL. Specifically, our proposed FedFV method can build a connection between fairness and conflicting gradients with large differences in the

Method	Parameters
AFL	$\eta_\lambda \in \{0.01, 0.1, 0.5\}$
qFedAvg	$q \in \{0.1, 0.2, 1, 2, 5, 15\}$
FedMGDA+	$\epsilon \in \{0.01, 0.05, 0.1, 0.5, 1\}$
FedFA	$(\alpha, \beta) \in \{(0.5, 0.5)\}, (\gamma_s, \gamma_c) \in \{(0.5, 0.9)\}$
FedFV	$\alpha \in \{0, 0.1, 0.2, \frac{1}{3}, 0.5, \frac{2}{3}\}, \tau \in \{0, 1, 3, 10\}$

Table 1: Method specific hyper-parameters.

magnitudes, and we prove its ability to mitigate two gradient conflicts (i.e., internal and external conflicts) and converge to Pareto stationary solutions.

5 Experiments

5.1 Experimental Setup

Datasets and Models

We evaluate FedFV on three public datasets: CIFAR-10 [Krizhevsky, 2012], Fashion MNIST [Xiao *et al.*, 2017] and MNIST [LeCun *et al.*, 1998]. We follow [McMahan *et al.*, 2017] to create non-I.I.D. datasets. For CIFAR-10, we sort all data records based on their classes, and then split them into 200 shards. We use 100 clients, and each client randomly picks 2 shards without replacement so that each has the same data size. We use 200 clients for MNIST and preprocess MNIST in the same way as CIFAR-10. The local dataset is split into training and testing data with percentages of 80% and 20%. For Fashion MNIST, we simply follow the setting in [Li *et al.*, 2019]. We use a feedforward neural network with 2 hidden layers on CIFAR-10 and Fashion MNIST. We use a CNN with 2 convolution layers on MNIST.

Baselines

We compare with the classical method FedAvg [McMahan *et al.*, 2017] and FL systems that address fairness in FL, including AFL [Mohri *et al.*, 2019], q-FedAvg [Li *et al.*, 2019], FedFa [Huang *et al.*, 2020] and FedMGDA+ [Hu *et al.*, 2020]. We compare with q-fedavg, FedFa and fedmgda+ on all three datasets and compare with AFL only on Fashion MNIST, because AFL is only suitable for small networks with dozens of clients.

Hyper-parameters

For all experiments, we fix the local epoch $E = 1$ and use batchsize $B_{CIFAR-10|MNIST} \in \{full, 64\}, B_{FashionMNIST} \in \{full, 400\}$ to run Stochastic Gradient Descent (SGD) on local datasets with stepsize $\eta \in \{0.01, 0.1\}$. We verify the different methods with hyper-parameters as listed in Table 1. We take the best performance of each method for the comparison.

Implementation

All our experiments are implemented on a 64g-MEM Ubuntu 16.04.6 server with 40 Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and 4 NVidia(R) 2080Ti GPUs. All code is implemented in PyTorch version 1.3.1. Please see https://github.com/WwZzz/easyFL for full details.

Method	Ave.	Var.	Worst 5%	Best 5%
FedAvg	46.85±0.65	12.57±1.50	19.84±6.55	69.28±1.17
qFedAvg$\vert_{q=0.1}$	47.02±0.89	13.16±1.84	18.72±6.94	70.16±2.06
qFedAvg$\vert_{q=0.2}$	46.91±0.90	13.09±1.84	18.88±7.00	70.16±2.10
qFedAvg$\vert_{q=1.0}$	46.79±0.73	11.72±1.00	22.80±3.39	68.00±1.60
qFedAvg$\vert_{q=2.0}$	46.36±0.38	10.85±0.76	24.64±2.17	66.80±2.02
qFedAvg$\vert_{q=5.0}$	45.25±0.42	**9.59±0.36**	26.56±1.03	63.60±1.13
FedFa	46.43±0.56	12.79±1.54	19.28±6.78	69.36±1.40
FedMGDA+$\vert_{\epsilon=0.01}$	45.65±0.21	10.94±0.87	25.12±2.34	67.44±1.20
FedMGDA+$\vert_{\epsilon=0.05}$	45.58±0.21	10.98±0.81	25.12±1.87	67.76±2.27
FedMGDA+$\vert_{\epsilon=0.1}$	45.52±0.17	11.32±0.86	24.32±2.24	68.48±2.68
FedMGDA+$\vert_{\epsilon=0.5}$	45.34±0.21	11.63±0.69	24.00±1.93	68.64±3.11
FedMGDA+$\vert_{\epsilon=1.0}$	45.34±0.22	11.64±0.66	24.00±1.93	68.64±3.11
FedFV$\vert_{\alpha=0.1,\tau=0}$	48.57±0.76	10.97±1.02	28.32±2.01	69.76±2.45
FedFV$\vert_{\alpha=0.2,\tau=0}$	48.54±0.64	10.56±0.96	28.88±1.92	69.20±2.01
FedFV$\vert_{\alpha=0.5,\tau=0}$	48.14±0.38	10.60±0.76	28.72±1.92	68.80±1.77
FedFV$\vert_{\alpha=0.1,\tau=1}$	49.34±0.56	10.74±0.89	28.56±2.54	70.24±1.49
FedFV$\vert_{\alpha=0.1,\tau=3}$	50.00±0.74	10.85±1.14	28.24±3.03	**70.96±1.00**
FedFV$\vert_{\alpha=0.1,\tau=10}$	**50.42±0.55**	9.70±0.96	**32.24±2.10**	69.68±2.84

Table 2: The average, the variance, the worst and the best of the test accuracy of all clients on CIFAR-10. All experiments are running over 2000 rounds with full batch size, learning rate $\eta = 0.1$ and local epochs $E = 1$. The reported results are averaged over 5 runs with different random seeds.

5.2 Experimental Results

Fairness

We first verify the advantage of FedFV in fairness. In Table 2, we list the the mean, variance, the worst 5% and the best 5% of test accuracy on 100 clients created by splitting CIFAR-10 where 10% of clients are sampled in each communication round. While reaching a higher mean of test accuracy across clients, FedFV also yields the lowest variance among all experiments except for q-FedAvg$|_{q=5.0}$. Note that q-FedAvg sacrifices the best performance of clients from 69.28% to 63.60% and causes a significant reduction of the mean of accuracy from 46.85% to 45.25%. When reaching a similar variance to q-FedAvg$|_{q=5.0}$ (9.59 versus 9.72), FedFV improves the mean of accuracy to 50.42%. Further, the worst 5% performance of all the experiments on FedFV is higher than that of any others, which indicates FedFV's effectiveness on protecting the worst performance.

Table 3 lists the model test accuracy on each client's data and the average and variance of the accuracy on Fashion-MNIST dataset. Different from settings in Table 2, we use a small network with only 3 clients, and the server selects all the clients to participate the training process in each communication round to eliminate the sample bias, which results in the absence of external conflicts. Therefore, we only use FedFV with $\alpha \in \{0, \frac{1}{3}, \frac{2}{3}\}, \tau = 0$ in this case. In our observation, even though FedFV with $\alpha = 0$ can still find the solution with the highest mean of test accuracy 80.48%, it's

Method	$shirt$	$pullover$	$T-shirt$	Ave.	Var.
FedAvg	64.26±1.07	87.03±1.40	89.97±1.30	80.42±0.76	11.50±0.95
AFL$\|_{\eta_\lambda=0.01}$	71.09±0.91	81.89±1.30	84.29±1.23	79.09±0.73	5.76±0.75
AFL$\|_{\eta_\lambda=0.1}$	76.34±0.63	79.00±1.04	79.06±1.21	78.13±0.72	1.27±0.66
AFL$\|_{\eta_\lambda=0.5}$	76.57±0.58	78.77±0.99	79.09±1.15	78.14±0.71	**1.12±0.61**
qFedAvg$\|_{q=5}$	71.29±0.85	81.46±1.08	82.86±0.97	78.53±0.64	5.16±0.69
qFedAvg$\|_{q=15}$	77.09±0.88	75.40±2.07	60.69±1.79	71.06±0.81	7.46±0.88
FedFa	63.80±1.10	86.86±1.28	89.97±1.30	80.21±0.73	11.68±0.93
FedMGDA+$\|_{\epsilon=0.05}$	44.63±1.63	57.77±5.72	**99.09±0.23**	67.16±1.65	23.39±0.57
FedMGDA+$\|_{\epsilon=0.1}$	72.26±3.12	79.71±4.47	86.03±4.27	79.33±1.11	6.45±2.21
FedMGDA+$\|_{\epsilon=1.0}$	72.46±3.32	79.74±4.48	85.66±4.88	79.29±1.14	6.42±2.23
FedFV$\|_{\alpha=0.0,\tau=0}$	61.06±1.21	**89.31±1.38**	91.06±1.29	**80.48±0.68**	13.76±1.02
FedFV$\|_{\alpha=\frac{1}{3},\tau=0}$	77.09±1.35	80.69±1.53	81.06±0.53	79.61±0.85	1.91±0.57
FedFV$\|_{\alpha=\frac{2}{3},\tau=0}$	**77.91±0.80**	81.46±1.99	81.46±1.51	80.28±1.09	1.77±0.87

Table 3: Test accuracy on the different clothing classes of Fashion MNIST dataset. All experiments are running over 200 rounds with full batch size, learning rate $\eta = 0.1$ and local epochs $E = 1$. The reported results are averaged over 5 runs with different random seeds.

	FedFV	FedFV$_{Random}$	FedFV$_{Reverse}$
$Var._{CIFAR10}$	**13.19±1.06**	14.12±0.55	16.28±0.72
$Var._{FMNIST}$	**13.76±1.02**	20.14±0.61	22.05±0.68

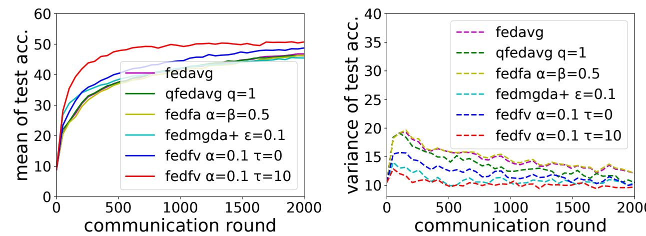
Table 4: The effects of projecting order to fairness. The results are averaged over 5 runs with different random seeds.

a unfair solution where the variance is up to 13.76. Instead, FedFV with $\alpha = \frac{2}{3}$ also finds a solution with high accuracy 80.28%, and simultaneously keep the variance in a low level. We notice that AFL$|_{\eta_\lambda=0.5}$ has the lowest variance. However, AFL only works for small works where the number of clients is limited. Despite of this, FedFV$|_{\alpha=\frac{2}{3}}$ outperforms AFL$|_{\eta_\lambda=0.5}$ on all the clients (i.e. shirt, pullover and T-shirt), as well as the mean of accuracy. Further, when compared to other methods except AFL, our method has the lowest variance down to 1.77, which validates the ability of FedFV to find a fair and accurate solution without the presence of sample bias.

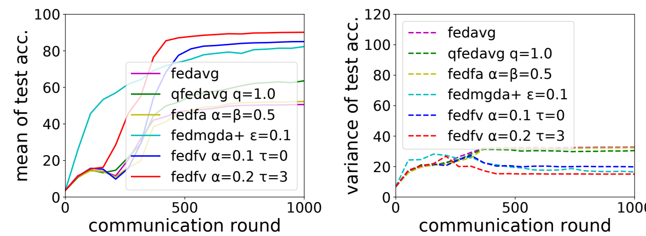
From Table 2 and Table 3, we observe that FedFV is able to ease the negative impact of conflicting gradients with largely different magnitudes, which results in a fairer and accurate model with lower variance.

Accuracy and Efficiency

We further show that FedFV outperforms existing works addressing fairness in terms of accuracy and efficiency on CIFAR-10 and MNIST. All the methods are tuned to their best performance. As shown in Figure 2, when we only mitigate the internal conflicts, which is done by FedFV($\tau = 0$), we already converge faster while keeping a lower variance than the others. In addition, when we mitigate both the internal and external conflicts by FedFV(τ>0), there is a substantial improvement in the accuracy and efficiency again. FedFV outperforms state-of-the-art methods by up to 7.2% on CIFAR-10 and 78% on MNIST, which verifies the ability of FedFV to reach a higher accuracy with less communication rounds while still keeping a low variance. Thus, we demonstrate that FedFV's advantages in saving communication cost and finding a better generalization.



(a) CIFAR-10



(b) MNIST

Figure 2: The mean (left) and the variance (right) of test accuracy on all clients on (a) CIFAR-10, and (b) MNIST. The results are averaged over 5 runs with different random seeds.

Effects of Projecting Order

To verify the effectiveness of our projecting order, we compare FedFV with another two cases: 1) projecting gradients in a random order of the projection targets; and 2) projecting gradients in an order of the projection targets that is reverse to FedFV, as shown in Table 4.

For CIFAR-10, 20% of clients are sampled in each communication round. For Fashion MNIST, all clients are selected in each communication round. We set $\alpha = 0, \tau = 0$ for all groups to confirm the effectiveness of the projecting order. If we project gradients to the targets in the loss-based order of FedFV, the variance is the lowest. Projecting gradients to the targets in a random order is also fairer than in an order reverse to FedFV, which indicates that the loss-based order used by FedFV helps improve fairness.

6 Conclusions and Future Work

We identified the issue of conflicting gradients with large differences in the magnitudes, which brings unfairness in FL. To address this issue, we propose the **Fed**erated **F**air a**V**eraging (**FedFV**) algorithm to mitigate the potential conflicts among clients before averaging their gradients. In addition, we show how well FedFV can mitigate the conflicts and how it converges to either a pareto stationary solution or the optimal on convex problems. Extensive experiments on a suite of federated datasets confirm that FedFV compares favorably against state-of-the-art methods in terms of fairness, accuracy and efficiency. In the future, we plan to build a complete theoretical analysis of FedFV on mitigating complex external conflicts.

A Proof

A.1 Proof of Theorem 1

Proof. For each gradient $g_i \in G$, we project g_i onto g_k's normal plane in an increasing order with k. Thus, we have update rules

$$\begin{cases} g_i^{(0)} = g_i, & k = 0 \\ g_i^{(k)} = g_i^{(k-1)} - \frac{g_i^{(k-1)} \cdot g_k}{||g_k||^2} g_k, & k = 1, 2, ..., m, k \neq i \\ g_i^{(k)} = g_i^{(k-1)}, & k = i \end{cases}$$

Therefore, $g_i^{(k)}$ never conflicts with g_k, since all potential conflicts have been eliminated by the updating rules. We then focus on how much the final gradient $\bar{g}'$ conflicts with each gradient in G. Firstly, the last in the projecting order is g_m, and the final gradient $\bar{g}'$ will not conflict with it. Then, we focus on the last but second gradient g_{m-1}. Following the update rules, we have

$$g_i^{(m)} = g_i^{(m-1)} - \frac{g_i^{(m-1)} \cdot g_m}{||g_m||^2} g_m \tag{2}$$

Let $\phi_{i,j}^{(k)}$ denote the angle between $g_i^{(k)}$ and g_j, and $\phi_{i,j} = \phi_{i,j}^{(0)}$, then the average of projected gradients is:

$$\begin{aligned} \bar{g}' &= \frac{1}{m} \sum_i^m g_i^{(m)} \\ &= \frac{1}{m} (\sum_{i \neq m}^m (g_i^{(m-1)} - \frac{g_i^{(m-1)} \cdot g_m}{||g_m||^2} g_m) + g_m^{(m-1)}) \\ &= \frac{1}{m} \sum_i^m g_i^{(m-1)} - \frac{1}{m} \sum_{i \neq m}^m ||g_i^{(m-1)}|| \cos \phi_{i,m}^{(m-1)} \frac{g_m}{||g_m||} \end{aligned} \tag{3}$$

Since $g_{m-1} \cdot \sum_i^m g_i^{(m-1)} \geq 0$,

$$\begin{aligned} g_{m-1} \cdot \bar{g}' &\geq g_{m-1} \cdot \frac{-1}{m} \sum_{i \neq m}^m ||g_i^{(m-1)}|| \cos \phi_{i,m}^{(m-1)} \frac{g_m}{||g_m||} \\ &= -\frac{1}{m} \sum_{i \neq m}^m ||g_i^{(m-1)}|| ||g_{m-1}|| \cos \phi_{i,m}^{(m-1)} \cos \phi_{m-1,m} \\ &\geq -\frac{\epsilon^2}{m} ||g_{m-1}|| \sum_{i \neq m}^m ||g_i^{(m-1)}|| \end{aligned} \tag{4}$$

Similarly, we can compute the conflicts between any gradient $g_k \in G$ and $\bar{g}'$ by removing the $g_i^{(k)}$ from $\bar{g}'$

$$\begin{aligned} g_k \cdot \bar{g}' &\geq g_k \cdot \sum_{j=k}^{m-1} (\frac{-1}{m} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \cos \phi_{i,j+1}^{(j)} \frac{g_{j+1}}{||g_{j+1}||}) \\ &= -\frac{||g_k||}{m} \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \cos \phi_{i,j+1}^{(j)} \cos \phi_{k,j+1} \\ &\geq -\frac{\epsilon^2}{m} ||g_k|| \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \\ |e_k \cdot \bar{g}'| &= |\frac{g_k}{||g_k||} \cdot \bar{g}'| \leq \frac{\epsilon^2}{m} \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \end{aligned} \tag{5}$$

Therefore, the later g_k serves as the projecting target of others, the smaller the upper bound of conflicts between $\bar{g}'$ and g_k is, since $\sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \leq \sum_{j=k-1}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}||$. $\square$

A.2 Proof of Theorem 2

Proof. According to (5), the upper bound of the conflict between any client gradient g_k and the final average gradient $\bar{g}'$ can be expressed as

$$|g_k \cdot \bar{g}'| \leq \frac{\epsilon_2^2}{m} ||g_k|| \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \tag{6}$$

With the update rules of FedFV, we can infer that

$$\begin{aligned} ||g_k^{(i)}||^2 &= ||g_k^{(i-1)} - ||g_k^{(i-1)}|| \cos \phi_{k,i}^{(i-1)} \frac{g_i}{||g_i||}||^2 \\ &= ||g_k^{(i-1)}||^2 - 2||g_k^{(i-1)}||^2 \cos^2 \phi_{k,i}^{(i-1)} + \\ &\quad ||g_k^{(i-1)}||^2 \cos^2 \phi_{k,i}^{(i-1)} \\ &= (1 - \cos^2 \phi_{k,i}^{(i-1)}) ||g_k^{(i-1)}||^2 \\ &\leq (1 - \epsilon_1^2) ||g_k^{(i-1)}||^2 \end{aligned} \tag{7}$$

Therefore, the maximum value of gradient conflict is bounded by

$$\begin{aligned} &|g_k \cdot \bar{g}'| \leq \frac{\epsilon_2^2}{m} ||g_k|| \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(j)}|| \\ &\leq \frac{\epsilon_2^2}{m} (\max_i ||g_i||) \sum_{j=k}^{m-1} \sum_{i \neq j+1}^m ||g_i^{(0)}|| (1 - \epsilon_1^2)^{\frac{j}{2}} \\ &\leq \frac{m-1}{m} \epsilon_2^2 (\max_i ||g_i||)^2 \sum_{j=k}^{m-1} (1 - \epsilon_1^2)^{\frac{j}{2}} \\ &= \frac{m-1}{m} (\max_i ||g_i||)^2 \frac{\epsilon_2^2 (1 - \epsilon_1^2)^{\frac{1}{2}} (1 - (1 - \epsilon_1^2)^{\frac{m-k}{2}})}{1 - (1 - \epsilon_1^2)^{\frac{1}{2}}} \\ &= \frac{m-1}{m} (\max_i ||g_i||)^2 f(m, k, \epsilon_1, \epsilon_2) \end{aligned} \tag{8}$$

$\square$

A.3 Proof of the Convergence

Proof of Theorem 3. Let g_1 and g_2 be the two types of users' updates of parameters in the tth communication round. When $F(\theta)$ is L-smooth, we have

$$F(\theta^{t+1}) \le F(\theta^t) + \nabla F(\theta)^T(\theta^{t+1} - \theta^t) + \frac{1}{2}L||\theta^{t+1} - \theta^t||_2^2 \tag{9}$$

There are two cases: conflicting gradients exist or otherwise. If there is no conflict between g_1 and g_2, which indicates $g_1 \cdot g_2 \ge 0$, FedFV updates as FedAvg does, simply computing an average of the two gradients and adding it to the parameters to obtain the new parameters $\theta^{t+1} = \theta^t - \eta\bar{g}$, where $g = \frac{1}{2}(g_1 + g_2)$. Therefore, when using step size $\eta \le \frac{1}{L}$, it will strictly decrease the objective function $F(\theta)$ [Li *et al.*, 2020c]. However, if $g_1 \cdot g_2 < 0$, FedFV will compute the update as:

$$\theta^{t+1} = \theta^t - \eta\bar{g}' = \theta^t - \eta(g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2}g_1 - \frac{g_1 \cdot g_2}{||g_2||^2}g_2) \tag{10}$$

Combining with (9), we have:

$$\begin{aligned}&F(\theta^{t+1}) \le F(\theta^t) - \eta(g_1+g_2)^T(g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2}g_1 - \\ &\frac{g_1 \cdot g_2}{||g_2||^2}g_2) + \frac{L}{2}\eta^2||g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2}g_1 - \frac{g_1 \cdot g_2}{||g_2||^2}g_2||_2^2 \\ &= F(\theta^t) - \eta||g_1 + g_1||^2 + \eta(g_1 + g_2)^T(\frac{g_1 \cdot g_2}{||g_1||^2}g_1 \\ &+ \frac{g_1 \cdot g_2}{||g_2||^2}g_2) + \frac{L}{2}\eta^2||g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2}g_1 - \frac{g_1 \cdot g_2}{||g_2||^2}g_2||_2^2\end{aligned} \tag{11}$$

Since $g_1 \cdot g_2 = ||g_1||||g_2||\cos\phi_{12}$, where ϕ_{12} denotes the angle between g_1 and g_2, after rearranging the items in this inequality, we have

$$\begin{aligned}F(\theta^{t+1}) \le F(\theta^t) - (\eta - \frac{L}{2}\eta^2)(1 - \cos^2\phi_{12})(||g_1||^2 + \\ ||g_2||^2) - Lt^2(1 - \cos^2\phi_{12})||g_1||||g_2||\cos\phi_{12}\end{aligned} \tag{12}$$

To decrease the objective function, the inequality below should be satisfied

$$\begin{aligned}-(\eta - \frac{L}{2}\eta^2)(1 - \cos^2\phi_{12})(||g_1||^2 + ||g_2||^2) - L\eta^2(1 - \\ \cos^2\phi_{12})||g_1||||g_2||\cos\phi_{12} \le 0\end{aligned} \tag{13}$$

$$(\eta - \frac{L}{2}\eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2}\eta^2 2||g_1||||g_2||\cos\phi_{12} \ge 0 \tag{14}$$

$$\begin{aligned}(\eta - L\eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2}\eta^2(||g_1||^2 + ||g_2||^2 + \\ 2||g_1||||g_2||\cos\phi_{12}) \ge 0\end{aligned} \tag{15}$$

$$(\eta - L\eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2}\eta^2||g_1 + g_2||^2 \ge 0 \tag{16}$$

With $\eta \le \frac{1}{L}$, we have $\eta - L\eta^2 \ge 0$, which promises the decrease of the objective function. On the other hand, we can infer that

$$-\eta(1 - \frac{L}{2}\eta) = \eta(\frac{L}{2}\eta - 1) \le \eta(\frac{1}{2} - 1) = -\frac{\eta}{2} \tag{17}$$

Combine (17) with (13), we have:

$$\begin{aligned}&F(\theta^{t+1}) \le F(\theta^t) - \frac{\eta}{2}(1 - \cos^2\phi_{12})(||g_1||^2 + ||g_2||^2) \\ &- \frac{\eta}{2}(1 - \cos^2\phi_{12})2||g_1||||g_2||\cos\phi_{12} \\ &= F(\theta_t) - \frac{\eta}{2}(1 - \cos^2\phi_{12})(||g_1||^2 + ||g_2||^2 + 2||g_1||||g_2|| \\ &\cos\phi_{12}) \\ &= F(\theta_t) - \frac{\eta}{2}(1 - \cos^2\phi_{12})||g_1 + g_2||^2 \\ &= F(\theta_t) - \frac{\eta}{2}(1 - \cos^2\phi_{12})||g||^2\end{aligned} \tag{18}$$

Therefore, the objective function will always degrade unless 1)$||g|| = 0$, which indicates it will reach the optimal θ^*, 2)$\cos\phi_{12} = -1$, then we can suppose $g_1 = -\beta g_2$ where $\beta > 0$, and we will get

$$\frac{1}{1+\beta}g_1 + \frac{\beta}{1+\beta}g_2 = 0 \tag{19}$$

So θ^t is a pareto stationary point as defined below:

Definition 4. *For smooth criteria $F_k(\theta)(1 \le k \le K)$, θ^0 is called a Pareto-stationary iff there exists some convex combination of the gradietns $\nabla F_k(\theta^0)$ that equals zero [Désidéri, 2012].*

□

Proof of Theorem 4. With the assumption that $F(\theta)$ is L-smooth, we can get the inequality:

$$F(\theta^{t+1}) \le F(\theta^t) + \nabla F(\theta)^T(\theta^{t+1} - \theta^t) + \frac{1}{2}L||\theta^{t+1} - \theta^t||_2^2. \tag{20}$$

Similar to the proof for Theorem 2, if there exist conflicts, then

$$\begin{aligned}F(\theta^{t+1}) &\le F(\theta^t) - \eta\bar{g} \cdot \bar{g}' + \frac{1}{2}L\eta^2||\bar{g}'||^2 \\ &\le F(\theta^t) - \frac{\eta}{2}||\bar{g}||||\bar{g}'|| + \frac{L}{2}\eta^2||\bar{g}'||^2 \\ &\le F(\theta^t) - \frac{\eta}{2}||\bar{g}||||\bar{g}'|| + \frac{L}{2}\eta^2||\bar{g}||||\bar{g}'|| \\ &\le F(\theta^t) + (-\frac{\eta}{2} + \frac{L}{2}\eta^2)||\bar{g}||||\bar{g}'||\end{aligned} \tag{21}$$

Since $(-\frac{\eta}{2} + \frac{L}{2}\eta^2) \le 0$ with $\eta \le \frac{1}{L}$, the average objective will always degrade if we repeatedly apply the update rules of FedFV unless 1)$||\bar{g}|| = 0$, which indicates it will finally reach the optimal θ^*, 2) $||\bar{g}'|| = 0$, which indicates that all pairs of conflicting gradients have a cosine similarity of -1, leading to an existence of convex combination p which satisfies $\sum p_i g_i = 0$, since we can easily choose a pair of conflicting gradients g_i and g_j as the proof of theorem 3 does and then set the weights of the rest zero. □

References

[Abay *et al.*, 2020] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv e-prints*, page arXiv:2012.02447, December 2020.

[Chaudhry *et al.*, 2018] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. *CoRR*, abs/1812.00420, 2018.

[Cho *et al.*, 2020] Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Yağan. Bandit-based communication-efficient client selection strategies for federated learning. *arXiv e-prints*, page arXiv:2012.08009, 2020.

[Désidéri, 2012] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313 – 318, 2012.

[Farajtabar *et al.*, 2019] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. *CoRR*, abs/1910.07104, 2019.

[Guo *et al.*, 2019] Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. Learning with long-term remembering: Following the lead of mixed stochastic gradient. *CoRR*, abs/1909.11763, 2019.

[Hu *et al.*, 2020] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Fedmgda+: Federated learning meets multi-objective optimization. *arXiv e-prints*, page arXiv:2006.11489, 2020.

[Huang *et al.*, 2020] Wei Huang, Tianrui Li, Dexian Wang, Shengdong Du, and Junbo Zhang. Fairness and accuracy in federated learning. *arXiv e-prints*, page arXiv:2012.10069, 2020.

[Kairouz *et al.*, 2019] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, and et al. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.

[Krizhevsky, 2012] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[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.*, 2019] Tian Li, Maziar Sanjabi, and Virginia Smith. Fair resource allocation in federated learning. *CoRR*, abs/1905.10497, 2019.

[Li *et al.*, 2020a] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Federated multi-task learning for competing constraints. *arXiv e-prints*, page arXiv:2012.04221, 2020.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv e-prints*, page arXiv:1812.06127, 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. *arXiv e-prints*, page arXiv:1907.02189, 2020.

[Lopez-Paz and Ranzato, 2017] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continuum learning. *CoRR*, abs/1706.08840, 2017.

[Lyu *et al.*, 2019] Lingjuan Lyu, Jiangshan Yu, Karthik Nandakumar, Yitong Li, Xingjun Ma, and Jiong Jin. Towards fair and decentralized privacy-preserving deep learning with blockchain. *CoRR*, abs/1906.01167, 2019.

[Lyu *et al.*, 2020] Lingjuan Lyu, Xinyi Xu, and Qian Wang. Collaborative fairness in federated learning. *arXiv e-prints*, page arXiv:2008.12161, 2020.

[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 Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. *CoRR*, abs/1902.00146, 2019.

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

[Xu and Lyu, 2020] Xinyi Xu and Lingjuan Lyu. Towards building a robust and fair federated learning system. *arXiv e-prints*, page arXiv:2011.10464, 2020.

[Yu *et al.*, 2020] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. *arXiv e-prints*, page arXiv:2001.06782, January 2020.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. *arXiv e-prints*, page arXiv:1806.00582, June 2018.