Federated Learning with Fair Averaging

Zheng Wang\textsuperscript{1}, Xiaoliang Fan\textsuperscript{1,*}, Jianzhong Qi\textsuperscript{2}, Chenglu Wen\textsuperscript{1}, Cheng Wang\textsuperscript{1} and Rongshan Yu\textsuperscript{1}

\textsuperscript{1}Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China
\textsuperscript{2}School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

\small{fuxiaoliang@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 -- conflict 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.

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 \textit{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 \textit{et al.}, 2019; Li \textit{et al.}, 2019; Mohri \textit{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 \textit{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 \textit{et al.} [2019] seek to balance the overall performance and fairness using a fair resource allocation method. Hu \textit{et al.} [2020] find a common degrade direction for all clients without sacrificing anyone’s benefit and demonstrate robustness to inflating loss attacks. Li \textit{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 conflict 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 suffer when being used by more senior users, as the younger generation may use mobile devices more frequently and contribute more training data. We argue that if projecting the gradients to mitigate the conflicts before averaging, the update direction $\hat{g}$ will be corrected, which is closer to the optimal and fairer.

![Figure 1: When there are conflicting gradients ($g_a \cdot g_b < 0, g_a \cdot g_c < 0$) and large differences in the gradient magnitudes, $\|g_a\| > \|g_b\| > \|g_c\|$, the average of the original gradients $\theta$ will be dominated by $g_a$, which will be far away from the global optimal $\theta^*$, 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 $\hat{g}$ will be corrected, which is closer to the optimal and fairer.](image)

\textsuperscript{*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 different 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 **FedFV** (Federated Fair averaging) 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 $\theta$ 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 $k$th 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., $F_k(\theta) = \frac{1}{n_k} \sum_{i \in D_k} l(\theta, \xi_i)$, where $D_k$ denotes the local dataset of the $k$th 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 $\bar{g}^t = \frac{1}{m} \sum_{k=1}^{m} p_k g_k^t$ with weight $p_k = \frac{1}{\sum_{k=1}^{m} n_k}$ in the $t$th 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 $g_i \in \mathbb{R}^d$ conflicts with client j’s gradient $g_j \in \mathbb{R}^d$ iff $g_i \cdot g_j < 0$.

### 3.1 FedFV

FedFV aims to mitigate the conflicts among clients before averaging their gradients. In the $t$th 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 $t$th 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

\textbf{Input:} $T$, $m$, $\alpha$, $\tau$, $\eta$, $\theta^0$, $p_k$, $k = 1, \ldots, K$,

1: Initialize $\theta_0$ and gradient history $GH = []$.
2: for $t = 0, 1, \ldots, T - 1$ do
3: Server samples a subset $S_t$ of $m$ clients with the prob. $p_k$ and sends the model $\theta^t$ to them.
4: Server receives the updates $g^t_k$ and the training loss $l^t_k$ from each client $k \in S_t$, where $g^t_k \leftarrow \theta^t - \theta^t_k$ and $\theta^t_k$ is updated by client $k$. Then server updates the trace $GH = \{g^t_1, g^t_2, \ldots, g^t_K\}$ of the latest updates of all clients.
5: Server sorts the clients’ updates into a project order list $PO_t = \{g^t_1, \ldots, g^t_m\}$ based on their losses, where $l^t_i \leq l^t_{i+1}$.
6: $g^t \leftarrow \text{MitigateInternalConflict}(PO_t, G_t, \alpha)$
7: if $t \geq \tau$ then
8: $g^t \leftarrow \text{MitigateExternalConflict}(g^t, GH, \tau)$
9: end if
10: $g^t_k = g^t/\|g^t\| \times \frac{1}{m} \sum_{i=1}^{m} g^t_i$ \quad $\|g^t\|$
11: Server updates the model $\theta^{t+1} \leftarrow \theta^t - g^t$
12: end for.

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.

\textbf{Theorem 1.} Suppose there is a set of gradients $G = \{g_1, g_2, \ldots, g_m\}$ where $g_i$ always conflicts with $g^{(t)}_j$ before projecting $g^{(t)}_j$ to $g_i$’s normal plane, and $g^{(t)}_j$ is obtained by projecting $g_i$ to the normal planes of other gradients in $G$ for $t_i$ times. Assuming that $|\cos g^{(t)}_j, g^{(t)}_j| \leq \epsilon, 0 < \epsilon \leq 1$, for each $g_i \in G$, as long as we iteratively project $g_i$ onto $g_j$’s normal plane (skipping $g_i$, it self) in the ascending order of $k$ where $k = 1, 2, \ldots, m$, the larger $k$ is, the smaller the upper bound of conflicts between the average $g^t = \frac{1}{m} \sum_{i=1}^{m} g^{(m)}_i$ and $g_k$ is.

\textbf{Proof.} See Appendix A.1.

Since projecting gradients favors 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 $t$th communication round to improve the model performance on the less well trained data. We allow $\alpha m$ clients to keep their original gradients to further enhance fairness. We will detail $\alpha$ 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 $t$th round can suffer a risk of being forgotten by the model whenever the combined update $g^t$ conflicts with its imaginary gradient $g^t_{imag}$. 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:

\textbf{Definition 3.} In the $t$th communication round, there are external conflicts if there is a client $h \notin S_t$ whose largest gradient $g^t_{h,k}$ 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 $\frac{1}{m} \sum_{k=1}^{m} \|g^t_k\|$ 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.
Theorem 3. Assume that there are only two types of users point according Theorems 3 and 4. mitigating conflicts and can be applied to seek for a balance.

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 distribut-
edly 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 distri-
bution 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 deep model-based communication-efficient client selection strategy to overcome bias. 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 norm 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 orthonormal 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

## Algorithm 3 MitigateExternalConflict

**Input:** $g^t, GH, \tau$

1. for round $t - i, i = \tau, \tau - 1, \ldots, 1$
2. $g_{con} \leftarrow 0$
3. for each client $k = 1, 2, \ldots, K$
4. if $t_k = t - i$
5. if $g^t \cdot g_k^t < 0$
6. $g_{con} \leftarrow g_{con} + g_k^t$
7. end if
8. end if
9. end for
10. if $g^t \cdot g_{con} < 0$
11. Server computes $g^t \leftarrow g^t - \frac{g^t \cdot g_{con}}{\|g_{con}\|^2} g_{con}$.
12. end if
13. end for
14. return $g^t$

## Theorem 2

Suppose that there is a set of gradients $G = \{g_1, g_2, \ldots, g_n\}$ where $g_i$ always conflicts with $g_j^{(t)}$ before projecting $g_j^{(t)}$ to $g_i$’s normal plane and $g_i$’s is obtained by projecting $g_i$ to the normal planes of different gradients in $G$ for $t_i$ times. If $\epsilon_1 \leq 1 + \cos g_i^{(t)}, g_j^{(t)} > 1 - \epsilon_2, 0 < \epsilon_1 \leq \epsilon_2 \leq 1$, then as long as we iteratively project $g_i$ onto $g_i$’s normal plane (skipping $g_i$ itself) in the ascending order of k where $k = 1, 2, \ldots, m$, the maximum value of $\|g_i \cdot g_j\|$ is bounded by $m^{-1}(\max_i \|g_i\|)^2 f(m, k, \epsilon_1, \epsilon_2)$, where $f(m, k, \epsilon_1, \epsilon_2)$ is obtained before $(1 - \epsilon_1^2) \frac{1}{1 - \epsilon_2^2}$.

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 $g_i$ and the original average gradient is $\max_k |g_i \cdot \bar{g}| \leq m^{-1} \epsilon_2 (\max_k \|g_i\|)^2$.

In practice, we mitigate all the conflicts that have been de-
tected to limit the upper bound of gradient conflicts of clients. We allow $\alpha m$ 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 $\alpha$ 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 step-size $\eta \leq \frac{1}{L}$ will converge to either 1) a pareto stationary point, 2) or the optimal $\theta^*$.

Proof. See Appendix A.3.
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-IID 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 pre-process 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 four-layer neural network with 3 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 $B_{\text{CIFAR-10,MNIST}} \in \{\text{full, 64}\}$, $B_{\text{Fashion,MNIST}} \in \{\text{full, 400}\}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFL</td>
<td>$\eta_\lambda \in {0.01, 0.1, 0.5}$</td>
</tr>
<tr>
<td>qFedAvg</td>
<td>$q \in {0.1, 0.2, 1, 2, 5, 15}$</td>
</tr>
<tr>
<td>FedMGDA+</td>
<td>$\epsilon \in {0.01, 0.05, 0.1, 0.5, 1}$</td>
</tr>
<tr>
<td>FedFA</td>
<td>$(\alpha, \beta) \in {(0.5, 0.5), (0.5, 0.9)}$</td>
</tr>
<tr>
<td>FedFV</td>
<td>$\alpha \in {0.0, 0.1, 0.2, 0.3, 0.5, 0.7}$, $\tau \in {0, 1, 3, 10}$</td>
</tr>
</tbody>
</table>

Table 1: Method specific hyper-parameters.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>shirt</th>
<th>pullover</th>
<th>$T - shirt$</th>
<th>Ave.</th>
<th>Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FedAvg</td>
<td>64.26±0.77</td>
<td>89.97±1.30</td>
<td>80.42±0.76</td>
<td>11.50±0.95</td>
<td></td>
</tr>
<tr>
<td>AFL</td>
<td>71.09±0.91</td>
<td>81.89±1.30</td>
<td>84.29±1.25</td>
<td>11.50±0.95</td>
<td></td>
</tr>
<tr>
<td>AFA</td>
<td>76.34±0.63</td>
<td>89.00±1.04</td>
<td>79.06±1.21</td>
<td>78.13±0.72</td>
<td>12.72±0.66</td>
</tr>
<tr>
<td>AFA+</td>
<td>76.57±0.58</td>
<td>87.78±0.99</td>
<td>79.09±1.15</td>
<td>78.14±0.71</td>
<td>11.22±0.61</td>
</tr>
<tr>
<td>qFedAvg</td>
<td>71.29±0.85</td>
<td>81.46±1.08</td>
<td>82.86±0.97</td>
<td>78.53±0.64</td>
<td>5.16±0.69</td>
</tr>
<tr>
<td>qFedAvg</td>
<td>15.00±0.80</td>
<td>75.40±2.07</td>
<td>60.69±1.79</td>
<td>71.06±0.81</td>
<td>7.46±0.88</td>
</tr>
<tr>
<td>FedFa</td>
<td>63.80±1.10</td>
<td>86.86±1.28</td>
<td>89.97±1.30</td>
<td>80.21±0.73</td>
<td>11.68±0.93</td>
</tr>
<tr>
<td>FedMGDA+</td>
<td>44.63±0.63</td>
<td>77.57±4.72</td>
<td>99.09±2.23</td>
<td>76.16±1.65</td>
<td>23.39±0.57</td>
</tr>
<tr>
<td>FedMGDA+</td>
<td>72.26±3.12</td>
<td>79.71±4.47</td>
<td>86.03±4.27</td>
<td>79.33±1.11</td>
<td>6.45±2.21</td>
</tr>
<tr>
<td>FedMGDA+</td>
<td>72.46±3.32</td>
<td>79.74±4.48</td>
<td>85.66±4.88</td>
<td>79.29±1.14</td>
<td>6.42±2.23</td>
</tr>
<tr>
<td>FedFV</td>
<td>61.06±1.21</td>
<td>89.31±1.38</td>
<td>81.06±1.20</td>
<td>80.48±0.68</td>
<td>13.76±1.02</td>
</tr>
<tr>
<td>FedFV+</td>
<td>77.09±1.35</td>
<td>80.69±1.53</td>
<td>81.06±0.53</td>
<td>79.61±0.85</td>
<td>1.91±0.57</td>
</tr>
<tr>
<td>FedFV</td>
<td>77.91±0.80</td>
<td>81.46±1.99</td>
<td>81.46±1.51</td>
<td>80.28±1.09</td>
<td>1.77±0.87</td>
</tr>
</tbody>
</table>

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

- **Var.CIFAR10**: $13.19±1.06$ for FedFV, $14.12±0.55$ for FedFV$_{Random}$, $16.28±0.72$ for FedFV$_{reverse}$
- **Var.MNIST**: $13.76±1.02$ for FedFV, $20.14±0.61$ for FedFV$_{Random}$, $22.05±0.68$ for FedFV$_{reverse}$

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.

while still keeping a low variance. Thus, we demonstrate that FedFV’s advantages in saving communication cost and finding a better generalization.

**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 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($\tau > 0$), 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.
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:

\[
\begin{aligned}
   g_i^{(0)} &= g_i, \\
g_i^{(k)} &= g_i^{(k-1)} - \frac{\langle g_i^{(k-1)}, g_k \rangle}{\|g_k\|^2} g_k, & k = 1, 2, \ldots, m, k \neq i \\
g_i^{(k)} &= g_i^{(k-1)}, & k = i
\end{aligned}
\]

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 have been eliminated by the updating rules. We then focus on the last but second gradient \( g_{m-1} \). Following the update rules, we have

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

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

\[
g_k \cdot \bar{g} \geq g_k \cdot \sum_{j=k}^{m-1} \frac{\langle g_i^{(j)}, \phi_{i,j+1} \rangle}{\|g_{j+1}\|} g_{j+1}
\]

\[
= -\frac{1}{m} \sum_{j=k}^{m-1} \sum_{i \neq j+1} \|g_i^{(j)}\| \cos \phi_{i,j,1} \cos \phi_{k,j+1}
\]

\[
\geq -\frac{\epsilon^2}{m} \sum_{j=k}^{m-1} \sum_{i \neq j+1} \|g_i^{(j)}\|
\]

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

\[
\sum_{j=k}^{m-1} \sum_{i \neq j+1} \|g_i^{(j)}\| \leq \sum_{j=k-1}^{m-1} \sum_{i \neq j+1} \|g_i^{(j)}\|.
\]

\[
\square
\]

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

\[
\|g_k \cdot \bar{g}\| \leq \frac{\epsilon^2}{m} \sum_{j=k}^{m-1} \sum_{i \neq j+1} \|g_i^{(j)}\|.
\]

With the update rules of FedAvg, we can infer that

\[
\|g_k\|^2 = \|g_k^{(i-1)} - \|g_k^{(i-1)} \cos \phi_{k,i} \|g_i\|\|^2
\]

\[
= \|g_k^{(i-1)}\|^2 - 2\|g_k^{(i-1)}\|^2 \cos^2 \phi_{k,i} + ||g_k^{(i-1)}\|^2 \cos^2 \phi_{k,i}
\]

\[
= (1 - \sum_{j=k}^{m} \cos \phi_{i,j}) \|g_k^{(i-1)}\|^2
\]

\[
\leq (1 - \epsilon_i) \|g_k^{(i-1)}\|^2
\]

Therefore, the maximum value of gradient conflict is bounded by

\[
|g_k \cdot \bar{g}| \leq \frac{\epsilon^2}{m} \sum_{i} \sum_{j=k}^{m-1} \|g_i^{(j)}\|
\]

\[
\leq \frac{\epsilon^2}{m} (\max_i \|g_i\|) \sum_{j=k}^{m-1} \sum_{i \neq j+1} \|g_i^{(0)}\| (1 - \epsilon_i) \frac{1}{2}
\]

\[
\leq \frac{m-1}{m} \epsilon_i^2 (\max_i \|g_i\|) \sum_{j=k}^{m-1} (1 - \epsilon_i) \frac{1}{2}
\]

\[
= \frac{m-1}{m} (\max_i \|g_i\|) \epsilon_i^2 (1 - \epsilon_i) \frac{1}{2} (1 - (1 - \epsilon_i)^{m-k})
\]

\[
= \frac{m-1}{m} (\max_i \|g_i\|)^2 f(m, k, \epsilon_1, \epsilon_2)
\]

\[
\square
\]
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 \( t \)-th communication round. When \( F(\theta) \) is \( L \)-smooth, we have

\[
F(\theta^{t+1}) \leq F(\theta^t) + \nabla F(\theta)^T (\theta^{t+1} - \theta^t) + \frac{L}{2} ||\theta^{t+1} - \theta^t||^2_g
\]

(9)

There are two cases: conflicting gradients exist or otherwise. If there is no conflict between \( g_1 \) and \( g_2 \), which indicates \( g_1 \cdot g_2 \geq 0 \), FedFV updates as FedAvg does, simply computing an average of the two gradients and adding it to the parameters to obtain the new parameters \( \theta^{t+1} = \theta^t - \eta \tilde{g} \), where \( g = \frac{1}{2}(g_1 + g_2) \). Therefore, when using step size \( \eta \leq \frac{1}{L} \), 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:

\[
\theta^{t+1} = \theta^t - \eta \tilde{g} = \theta^t - \eta(g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2} g_1 - \frac{g_1 \cdot g_2}{||g_2||^2} g_2)
\]

Combining with (9), we have:

\[
F(\theta^{t+1}) \leq F(\theta^t) - \eta ||g_1 + g_2||^2 \cdot \eta (g_1 + g_2 - \frac{g_1 \cdot g_2}{||g_1||^2} g_1 - \frac{g_1 \cdot g_2}{||g_2||^2} g_2)
\]

(10)

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

\[
F(\theta^{t+1}) \leq F(\theta^t) - \eta (1 - \cos^2 \phi_{12})(||g_1||^2 + ||g_2||^2) - L\eta^2(1 - \cos^2 \phi_{12})||g_1|| ||g_2|| \cos \phi_{12} \leq 0
\]

(12)

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

\[
(\eta - \frac{L}{2} \eta^2)(1 - \cos^2 \phi_{12})(||g_1||^2 + ||g_2||^2) - L\eta^2(1 - \cos^2 \phi_{12})||g_1|| ||g_2|| \cos \phi_{12} \leq 0
\]

(13)

\[
(\eta - \frac{L}{2} \eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2} \eta^2 ||g_1|| ||g_2|| \cos \phi_{12} \geq 0
\]

(14)

\[
(\eta - L\eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2} \eta^2(2||g_1||^2 + ||g_2||^2 + 2||g_1|| ||g_2|| \cos \phi_{12}) \geq 0
\]

(15)

\[
(\eta - L\eta^2)(||g_1||^2 + ||g_2||^2) + \frac{L}{2} \eta^2 ||g_1|| + ||g_2||^2 \geq 0
\]

(16)

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

\[
-\eta(1 - \frac{L}{2} \eta) = -\eta(\frac{1}{2} - 1) = -\frac{\eta}{2}
\]

(17)

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

\[
F(\theta^{t+1}) \leq F(\theta^t) - \frac{\eta}{2}(1 - \cos^2 \phi_{12})||g_1||^2 + ||g_2||^2
\]

\[
- \frac{\eta}{2}(1 - \cos^2 \phi_{12})||g_1|| ||g_2|| \cos \phi_{12}
\]

\[
= F(\theta^t) - \frac{\eta}{2}(1 - \cos^2 \phi_{12})||g_1||^2 + ||g_2||^2 + 2||g_1|| ||g_2|| \cos \phi_{12}
\]

(18)

Therefore, the objective function will always degrade unless \( 1)||g|| = 0 \), which indicates it will reach the optimal \( \theta^* \). Therefore, the average objective will always degrade unless \( \frac{\eta}{2}(1 - \cos^2 \phi_{12})||g_1|| + ||g_2||^2 \)

\[
= F(\theta^t) - \frac{\eta}{2}(1 - \cos^2 \phi_{12})||g||^2
\]

(19)

So \( \theta^t \) is a pareto stationary point as defined below:

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

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

\[
F(\theta^{t+1}) \leq F(\theta^t) + \nabla F(\theta)^T (\theta^{t+1} - \theta^t) + \frac{L}{2} ||\theta^{t+1} - \theta^t||^2
\]

(20)

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

\[
F(\theta^{t+1}) \leq F(\theta^t) - \eta \tilde{g} + \frac{1}{2} L \eta^2 ||\tilde{g}||^2
\]

(21)

Since \( (-\eta + \frac{L}{2} \eta^2)^2 \leq 0 \) with \( \eta \leq \frac{1}{L} \), the average objective will always degrade if we repeatedly apply the update rules of FedFV unless \( ||\tilde{g}|| = 0 \), which indicates it will finally reach the optimal \( \theta^* \). Therefore, the objective will always degrade unless the stepsize is chosen so 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


