

CSAFL: A Clustered Semi-Asynchronous Federated Learning Framework

Yu Zhang*, Moming Duan*, Duo Liu*, Li Li*, Ao Ren*, Xianzhang Chen*, Yujuan Tan*, Chengliang Wang*

*College of Computer Science, Chongqing University, Chongqing, China

Email: zhangyucqu9@gmail.com, {duanmoming, liuduo}@cqu.edu.cn

Abstract—Federated learning (FL) is an emerging distributed machine learning paradigm that protects privacy and tackles the problem of isolated data islands. At present, there are two main communication strategies of FL: synchronous FL and asynchronous FL. The advantages of synchronous FL are that the model has high precision and fast convergence speed. However, this synchronous communication strategy has the risk that the central server waits too long for the devices, namely, the straggler effect which has a negative impact on some time-critical applications. Asynchronous FL has a natural advantage in mitigating the straggler effect, but there are threats of model quality degradation and server crash. Therefore, we combine the advantages of these two strategies to propose a clustered semi-asynchronous federated learning (CSAFL) framework. We evaluate CSAFL based on four imbalanced federated datasets in a non-IID setting and compare CSAFL to the baseline methods. The experimental results show that CSAFL significantly improves test accuracy by more than +5% on the four datasets compared to *TA-FedAvg*. In particular, CSAFL improves absolute test accuracy by +34.4% on non-IID FEMNIST compared to *TA-FedAvg*.

I. INTRODUCTION

Federated learning (FL) [1]–[3] is an emerging machine learning framework that utilizes multiple edge devices to jointly train a global model under the coordination of the central server. The training process of FL is divided into plenty of communication rounds. In each communication round, the edge devices download and use the global model parameters from the central server to perform optimization with local data for obtaining the local model parameters. Finally, each edge device updates the local model parameters to the central server for aggregating new global model parameters. Throughout the training process of the federated network, the training data of the client is always kept locally and is not transmitted, which protects data privacy [4]–[6]. In practice, The FL framework plays a crucial role in supporting privacy-sensitive applications on edge devices [7].

There are two model updating mechanisms for federated learning: synchronous and asynchronous. As shown in Fig. 1, we observe the training process of the two update strategies. Due to the different update mechanisms, these two FL frameworks have their own merits and drawbacks:

- Synchronous FL: In the synchronous FL, all clients need to download the global model parameters at a unified time node, and the central server waits for all clients to complete the training tasks. The merits of this synchronous updating strategy are the model has fast convergence

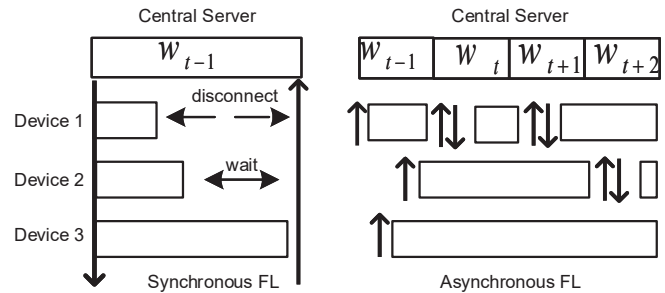


Fig. 1. The training procedures of synchronous FL and asynchronous FL

and high precision. The only drawback is the straggler effect [8] caused by network resources or poor hardware resources, which puts the server fall into an idle state.

- Asynchronous FL: Under the asynchronous FL, the central server cooperates with clients that complete the training tasks, rather than waiting for all clients, and each client can ask the central server for the new global model parameters when the training procedure is completed. Although this model updating strategy avoids the server falling into the idle state, it also has some shortcomings. On the one hand, this strategy makes the data transmission larger, which may cause the server to crash [9]. On the other hand, the gradient divergence caused by asynchronous updating will further degrade the performance of the model.

There have been many efforts in FL synchronous or asynchronous training algorithms. McMahan *et al.* propose the FL framework *Federated Averaging (FedAvg)* [1] and Li *et al.* experimentally and theoretically prove that *FedAvg* can achieve $\mathcal{O}(1/T)$ convergence rate with decayed learning rate under a statistical heterogeneous setting [10]. To relieve the effect of imbalanced data on accuracy, Duan *et al.* propose a self-balancing FL framework *Astraea* [11]. *FedAvg* and *Astraea* are both synchronous FL frameworks. However, these algorithms do not consider the stragglers, which is unfriendly for time-critical applications. Asynchronous FL framework mitigate the straggler effect. Xie *et al.* Proposed *FedAsync* [12] framework with a staleness function. Chen *et al.* Propose an asynchronous algorithm to tackle the challenges associated with heterogeneous edge devices, such as computational loads, stragglers [13]. However, these frameworks ignore the situa-

tion that the server may crash due to the continuous processing of requests from the clients [9].

In this paper, we propose a novel clustered semi-asynchronous federated learning (CSAFL) framework, which mitigates the straggler effect and controls model staleness in the asynchronous update for accuracy improvement. CSAFL leverages a spectral clustering algorithm [14], which groups clients according to the affinity matrix constructed by the delay and direction of clients' model update. In each communication round with fixed time budget, the selected clients do synchronous or asynchronous update independently, and contribute the update parameters to the group model they belong to. The main contributions of this paper are summarized as follows:

- We propose a novel clustered semi-asynchronous federated learning (CSAFL) framework. As far as I know, this is the first framework that combines synchronous and asynchronous update mechanisms.
- We design two strategies to alleviate the model staleness problem caused by asynchronous updating. The first is to leverage a spectral clustering algorithm to divide clients with different learning objectives into multiple groups. The second is by limiting the model delay.
- We evaluate CSAFL on four federated datasets, and show its effectiveness in mitigating the straggler effect compared to *T-Fedavg*, and show the accuracy improvement of the model is more than +5% on the four datasets compared to *TA-FedAvg*. Specially, the maximum accuracy improvement is +34.4% on FFEMNIST dataset compared to *TA-FedAvg*.

The rest of this paper is organized as follows. Section II outlines the background of synchronous FL and asynchronous FL. Section III shows the motivation of the CSAFL framework. Section IV details the design of the CSAFL framework. Evaluation results are presented and analyzed in Section V. Section VI concludes the paper.

II. BACKGROUND AND RELATED WORK

A. Federated Learning

In this section, we first present the vanilla synchronous FL framework *FedAvg*. Then we briefly introduce the asynchronous FL framework based on *Fedavg*.

McMahan *et al.* first propose the concept of federated learning, and later propose a widely used federated learning framework *FedAvg* [1], which involves solving a machine learning problem by loosely combining multiple devices under the coordination of a central server. Unlike the traditional distributed machine learning method, the computing nodes in the FL framework keep the training data locally and do not exchange or transfer the training data in the federated network. Instead, only the local updates of each computing node are transmitted, which reduces privacy risks. Nevertheless, the central server has no control over computing nodes. For example, computing nodes can join or drop out of the federated network at any time. In particular, the optimization goal of

Algorithm 1 Federated Averaging

```

1: procedure FL SERVER TRAINING
2:   Initialize global model  $\mathbf{w}_0$ .
3:   for each communication round  $t = 1, 2, \dots, T$  do
4:      $S_t \leftarrow$  Randomly select  $K$  clients from  $N$  clients.
5:     Server broadcasts  $\mathbf{w}_t$  to the selected clients.
6:     for each client  $c_i \in S_t$  parallelly do
7:        $\mathbf{w}_{t+1}^i \leftarrow \text{ClientUpdate}(i, \mathbf{w}_t)$ .
8:      $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \sum_{c_i \in S_t} \frac{n_i}{n} \mathbf{w}_{t+1}^i$ 

9: function CLIENTUPDATE( $i, \mathbf{w}$ )
10:   $\hat{\mathbf{w}} \leftarrow \mathbf{w}$ 
11:  for each local epoch  $e = 1, 2, \dots, E$  do
12:    for each local batch  $b \in \mathcal{B}$  do
13:       $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla L(b; \mathbf{w})$ 
14:  return  $\mathbf{w}$ 

```

general FL is:

$$\min_{\mathbf{w}} f(\mathbf{w}) \triangleq \sum_{k=1}^N p_k F_k(\mathbf{w}), \quad (1)$$

which means solving model updates \mathbf{w} when the value of $f(\mathbf{w})$ is the smallest. In (1), N is the number of clients participating in training, p_k refers to the participation weight of the k -th device, and $p_k \geq 0$, $\sum_k p_k = 1$, $F_k(\mathbf{w})$ is the local objective optimization function of the k -th device, we define $F_k(\mathbf{w})$ as:

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{(x_i, y_i) \sim p_{data}^k} l(x_i, y_i, \mathbf{w}), \quad (2)$$

where n_k is the training data size of the k -th client, (x_i, y_i) is the sample of the k -th client and obeys the data distribution p_{data} . $l(x_i, y_i, \mathbf{w})$ is the prediction loss function on (x_i, y_i) .

Specifically, the synchronous FL based on *FedAvg* mainly includes a central server maintaining the global model and multiple participants, which communicate through network connections with the server. In each communication round t , the server first selects a part of the client K from all clients N to participate in the training tasks and then broadcasts the global model \mathbf{w}_t through the network into the selected clients. The client c_i updates the \mathbf{w}_{t+1}^i locally based on \mathbf{w}_t . When the client completes the local training procedure, it transmits \mathbf{w}_{t+1}^i to the server through the network. The server waits for all clients to complete the training task. Finally, the server aggregates the new global model \mathbf{w}_{t+1} by averaging the local solutions of the clients. For an FL task, to achieve the target accuracy, it usually needs hundreds of communication rounds.

The pseudo-code of *FedAvg* is shown in Algorithm 1, where S_t is a random subset of K clients randomly selected in each communication round. Minibatch \mathcal{B} is the batch number of training data, and b refers to the subset of training data separated by \mathcal{B} . Local epoch E is the number of local training. n_i is the data size of client c_i and $n = \sum_{c_i \in S_t} n_i$ is the total

data size of selected clients. η is the learning rate of the local solver.

As shown in Algorithm 1, we note that the server needs to wait for all clients to complete the synchronous FL system's training tasks, which is negative for some time-critical applications. Therefore, the asynchronous FL framework is proposed. We briefly introduce the asynchronous FL framework based on *FedAvg*, where each client updates the local model to the global model independently. Whenever the server receives the local update from the client, it will refresh the global model. Therefore, the server does not need to wait for stragglers for aggregation.

B. Related Work

Federated Learning (FL), first proposed by Google, is a new approach to fitting machine learning into the edge. Existing FL frameworks can be classified into synchronous FL and asynchronous FL according to model updating. Most of the studies on synchronous FL do not consider the issues of stragglers due to the device heterogeneity and the instability of network conditions [1] [15] [16]. To eliminate the straggler effect on the statistical heterogeneity, Li *et al.* [17] propose a near term, experimentally prove this term can improve the stability of the framework and provide convergence guarantees in theory. However, the server still needs to wait for the local updates of stragglers before aggregation. In order to address the issue that all clients have to wait for the slowest one, Li *et al.* [18] propose a strategy to let the central server only accept the responses of the top threshold k clients, and the rest of the clients are regarded as stragglers. This method simply dropouts the stragglers, ignores the possibility of valuable data on the stragglers. In addition, to mitigate the impact of stragglers, Dhakal *et al.* [19] develop a coded computing technique for FL where sever compensates the gradient of the stragglers. However, this compensation is the result of the sever calculates the gradients based on the client's parity data. There is a risk of privacy leakage.

Asynchronous FL has a natural advantage over synchronous FL in solving the straggler effect [20] [21] [22], where the server can aggregate without waiting for stragglers. Chen *et al.* propose a *ASO-Fed* framework [13], which updates the global model in an asynchronous manner to tackle stragglers. However, gradient staleness is not considered, which may threaten the convergence of the model. Especially, Xie *et al.* develop a *FedAsync* [12] algorithm which combines a function of staleness with asynchronous update protocol. However, the clients will continue to transmit a large amount of data to the server, which may cause the server to crash. In terms of reducing data transmission, Wu *et al.* propose a *SAFA* [20] protocol which divides clients into three types, in which asynchronous clients continuously perform local update until the difference between local update version and global model version reaches tolerance. Although *SAFA* considers model staleness, the server needs to wait for the asynchronous clients.

TABLE I
THE STATISTICS OF THE FOUR FEDERATED DATASETS

Task	Dataset	Model	devices	samples
Image Classification	MNIST	MCLR	1000	69035
	FEMNIST		200	18345
Simulation Test	Synthetic(0.8,0.5)		100	75349
Sentiment Analysis	Sent140	LSTM	772	40783

III. PRELIMINARY CASE STUDIES AND MOTIVATION

In this section, we illustrate our preliminary case studies, which guide the motivation for a new grouping model update strategy in FL.

Asynchronous FL has a natural advantage in solving the straggler effect in FL [20], but the model delay caused by asynchronous update may affect the accuracy of the model. To investigate the influence of different updating strategies on the accuracy of the model influence of different model updating strategies, we study the training process of synchronous FL and asynchronous FL. More specifically, we set up this test based on a 62- class Federated Extended MNIST [23] (FEMNIST) dataset using the multinomial logistic regression model. The statistical information of FEMNIST dataset is shown in Table I. In order to show the straggler effect clearly, we set a fixed time budget hyper-parameter for each communication round. So in a communication round, the client may do multiple synchronous or asynchronous updates. Under a fixed time budget, we set up two groups of experiments based on *FedAvg*, which are *T-FedAvg* and *TA-FedAvg*. *TA-FedAvg* is an asynchronous update algorithm based on *FedAvg*. In this FL test, we adopt the same notation as [1]: the number of all clients $K = 200$, the size of local minibatch $B = 10$, the number of local epochs $E = 10$, the number of round $T = 200$, the learning rate $\eta = 0.03$, the number of selected clients per round is 20. We add the time budget of each round H . In this test, $H = 15000ms$.

The top-1 test accuracy of the two model update strategies based on FEMNIST dataset is shown in Fig. 2. For the *TA-FedAvg* algorithm, a 26.3% reduction in the top-1 test accuracy compared to the *T-FedAvg* algorithm. This test shows that the asynchronous update strategy harms the accuracy of the model. In the case of synchronous model update, all clients simultaneously download the global model parameters from the server. There is no client training with old model

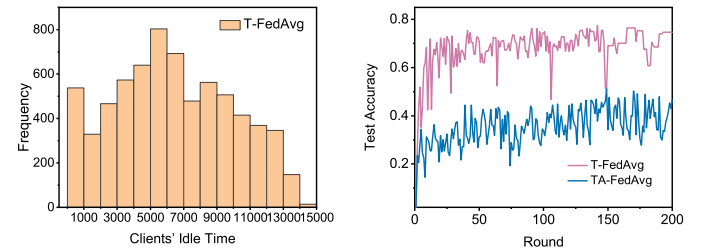


Fig. 2. The straggler effect in the synchronous FL framework can be observed in the figure on the left, and the frequency distribution of the clients' idle time is unbalanced; The figure on the right shows the accuracy comparison of asynchronous FL and synchronous FL in the process of model training.

parameters in the same communication round. Obviously, the model delay of each client is 1. However, in the case of asynchronous model update, When the server receives the local model parameters uploaded by one client, the global model on the server has been updated many times by other clients. Therefore, we realize that the model delay of each client is not steady. Due to the instability of model delay, the accuracy of asynchronous updates has decreased significantly.

As shown in Fig. 2, we can observe the clients' idle time distribution, which clearly shows that the server needs to wait too long, and the idle time is unbalanced in synchronization FL. However, in asynchronous FL, the server does not need to wait for all clients to complete the training tasks of each round. That is, when a client finishes uploading local model parameters, the server immediately refresh the global model, so the idle time of each client is 0. We assume that the communication time for the client to download model parameters from the server is negligible.

The preliminary case study shows that there is a straggler effect in the synchronous model update. In addition, the asynchronous model update strategy has the effect of precision degradation caused by model delay. We are also inspired by the Iterative Federated Clustering Algorithm (IFCA) framework proposed by Avishek [24], which considers different groups of users have their objectives. Therefore, we propose a novel clustered semi-asynchronous federated learning (CSAFL) framework.

IV. CSAFL

A. Framework Overview

To tackle the straggler effect in synchronous FL and model staleness in asynchronous FL, we propose a novel clustered semi-asynchronous federated learning (CSAFL) framework, which leverages a clustering algorithm based on similarity metrics to group clients. CSAFL combines synchronous FL and asynchronous FL to drive clients to update local model parameters to the same group's group model. As far as we know, this is the first paper to combine synchronous and asynchronous update mechanisms.

Our model architecture is shown in Fig. 3, CSAFL includes several groups, which can be deployed to the central server, or some devices in the middle layer, such as the edge server. They maintain the group models. In a group, the group model is the global model and the latest model. In this paper, we assume that all groups are deployed on a central server. CSAFL also contains a number of clients, which can be mobile phones, IoT devices, etc. There is a one-to-one communication between the client and its group. Due to the asynchronous update mechanism, the model maintained on the client is not necessarily the latest.

The training process of CSAFL is shown in Algorithm 2, CSAFL first initializes the group model W_0^G . In each communication round, the selected clients update the group model in its group, as shown in Fig. 3. For group g_x , clients A, B, C are the selected clients in one communication round.

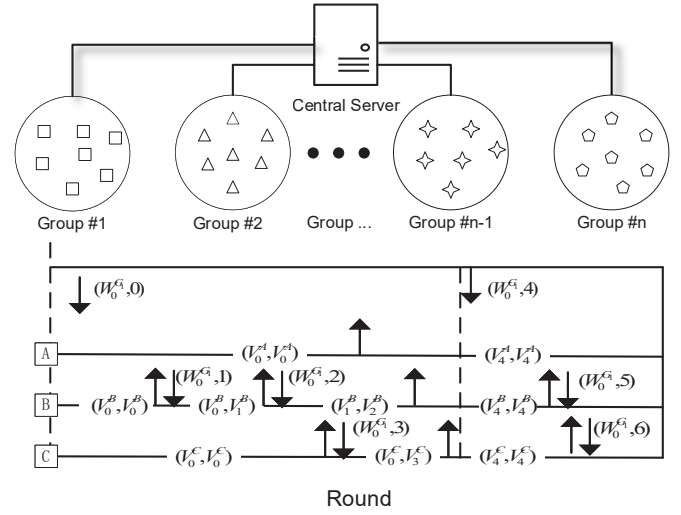


Fig. 3. The framework of CSAFL. In a group training process, group 1 randomly selects three clients, A, B, C. The up arrow indicates that the client updates the local model to the group model, and the down arrow indicates that the client downloads the parameters of the group model

The update process within the group g_x is divided into the following five steps:

- 1) Group g_x broadcasts the initial model to clients A, B, C and passes the version number of the group model, which is represented by $(W^{g_x}, 0)$
- 2) After the clients A, B, C receive the latest group model, they independently update the version number of the local model, such as, for client A (V_{pre}^A, V_{new}^A)
- 3) Clients A, B, C asynchronously update the group model according to its computing capacities and communication conditions. Each process of the asynchronous update will increase the version number of the group model. Before update operation, the difference between V_{pre} and V_{new} is calculated. If the difference is more significant than the tolerance H of gradient staleness, the clients whose time budgets are not exhausted are forced to update synchronously.
- 4) After each synchronization update, all clients download the same group model.
- 5) Repeat step 2 to step 4 until the time budget of the communication round is exhausted.

Similarly, when other clients are selected, the training process within the group is the same. Each group completes its own intra group update without interference. Each group has the same number of communication rounds.

We cover more details of CSAFL in the following four sections.

B. Clustering Strategy

Clustering Algorithm

As shown in Algorithm 2, we apply spectral clustering in this paper. Spectral clustering is more suitable for our framework, the following reasons:

- It is more general than the K-Means clustering algorithm [25], concretely, it is suitable for both convex sets and non-convex sets.
- It is not very sensitive to stragglers.
- It uses the Laplacian Eigenmaps for high-dimensional data, which reduces the load of calculation.

Similarity Metrics

The traditional clustering federated learning (CFL) framework is based on the similarity of sub-optimization tasks, which only considers the similarity of gradient direction [26] [27]. But in FL, this is not comprehensive. For example, In the asynchronous FL, we assume that the network conditions are same, and the transmission capacities of the devices are also fixed, and the device c_A with high computing power and the device c_B with poor computing power are divided into a group, the c_A will constantly refresh the global model with its own local updates, and the result of model training is more inclined to the c_A . In the synchronous FL, this grouping strategy will lead to the model training inefficient. Similarly, the data amount of different devices and different network conditions will make this grouping strategy have a bad impact on the performance of the model. Therefore, the data amount and the computing power of the device, and the network condition between the devices and the server may also be indicators of devices' similarity. The above three indicators are represented by the local model update latency of the client, so we determine the metrics for measuring the clients' similarity: gradient direction and the latency of model update. In the local update of a client, the local update latency can be divided into two parts: computation latency and communication latency, and we define the following three delay models:

1) *Computation Latency*: To quantify the randomness of the client's computing capabilities, we use a shift exponential distribution [28]:

$$P[t_{cp}^i < t] = \begin{cases} 1 - e^{-\frac{\mu_i}{d_i}(t-x_i d_i)} & t \geq x_i d_i \\ 0 & otherwise \end{cases}, \quad (3)$$

where $x_i > 0$ and $\mu_i > 0$ are parameters related to the computing capabilities of the client c_i . $x_i > 0$ and $\mu_i > 0$ represent the maximum and fluctuation of the client c_i computation capabilities, respectively. d_i is the size of the dataset on the client c_i , and t is the computation latency of the client c_i . Because the physical facility's computing capabilities where the group is located is much greater than that of the client, We ignore the delay due to group model aggregation. Based on (3), we can get the mathematical expectation of computation latency for client c_i as follows:

$$t_{ecp}^i = x_i d_i + \frac{d_i}{\mu_i}, \quad (4)$$

2) *Communication Latency*: We consider such a communication scenario, where all the clients participating in the training are within the cell radius of the central server. Those who are out of range can not participate in training. There is path loss in the transmission link. Given an FDMA system

Algorithm 2 clustered semi-asynchronous federated learning

Input: number of all clients R , number of selected clients per round K , set of groups G , learning rate η , local minibatch size \mathcal{B} , number of local epochs E , global initial model w_0 , time budget of each communication round H , number of round T , delay threshold per round L .

Output: updated group model parameters \mathcal{W}_T^G .

```

1: procedure FEDSPECASYNC TRAINING
2:   Initialize group model.
3:   for all  $g_i \in G$ , initial  $w_0^{g_i}$  to  $w_0$ 
4:    $\mathcal{W}_0^G \leftarrow [w_0^{g_1}, \dots, w_0^{g_n}]$ .
5:   for each communication round  $t = 1, 2, \dots, T$  do
6:      $S_t \leftarrow$  Server selects random  $K$  clients from all clients  $R$ .
7:     for all  $g_i \in G$ , initial  $v_{g_i}$  to 0,  $S_t^{g_i} \leftarrow \{c_j \mid c_j \in g_i \cdot clients, \forall c_j \in S_t\}$ .
8:     for each group  $g_i$  in  $G$  parallelly do
9:       for all  $c_j \in S_t^{g_i}$ , initial  $v_{pre}^{c_j}$  to 0,  $v_{new}^{c_j}$  to 0,  $w_{last}^{c_j}$  to  $w_0^{g_i}$ 
10:      for each client  $c_j$  in  $S_t^{g_i}$  parallelly do
11:        while  $H$  is not exhausted do
12:           $\Delta v \leftarrow v_{new}^{c_j} - v_{pre}^{c_j}$ .
13:          if  $\Delta v > L$  then
14:             $S_{syn} \leftarrow \{c_j \mid \text{the } H \text{ of } c_j \text{ is not exhausted, } \forall c_j \in S_t^{g_i}\}$ .
15:            SynUpdates( $S_{syn}$ )
16:          else
17:             $w_{last}^{c_j} \leftarrow \text{ModelUpdates}(c_j, w_{last}^{c_j})$ .
18:             $w_T^{g_i} \leftarrow w_{last}^{c_j}$ .
19:             $v_{g_i} \leftarrow v_{g_i} + 1$ .
20:             $v_{prev}^{c_j} \leftarrow v_{new}^{c_j}$ .
21:             $v_{new}^{c_j} \leftarrow v_{g_i}$ .
22: function MODELUPDATES( $c, w$ )
23:    $\mathcal{B} \leftarrow$  split dataset of client  $c$  into batches of size  $B$ .
24:   for each local epoch  $e$  from 1 to  $E$  do
25:     for each local batch  $b \in \mathcal{B}$  do
26:        $w \leftarrow w - \eta \nabla L(b; w)$ .
27:   return  $w$ .
28: function SYNUPDATES( $S_{syn}$ )
29:   for each client  $c_z$  in  $S_{syn}$  parallelly do
30:      $w_{last}^{c_z} \leftarrow \text{ModelUpdates}(c_z, w_{last}^{c_z})$ 
31:      $v_{g_i} \leftarrow v_{g_i} + 1$ .
32:      $n \leftarrow \sum_{c_z \in S_{syn}} n^{c_z}$ .
33:      $w_{last}^{g_i} \leftarrow \sum_{c_z \in S_{syn}} \frac{n_i}{n} w_{last}^{c_z}$ .
34:     for each client  $c_z$  in  $S_{syn}$  parallelly do
35:        $w_{last}^{c_z} \leftarrow w_{last}^{g_i}$ .
36:        $v_{prev}^{c_z} \leftarrow v_{g_i}$ .
37:        $v_{new}^{c_z} \leftarrow v_{g_i}$ .
38:        $w_T^{g_i} \leftarrow w_{last}^{c_z}$ .

```

Algorithm 3 Grouping Clients

Input: number of all clients K , learning rate η , local mini-batch size B , number of local epochs E , global initial model w_0 , number of group n pre-training hyper-parameter α , weight of time β .

Output: set of groups G

```

1: procedure GROUPING CLIENTS
2:    $\mathcal{M} \leftarrow \text{Calculate } \mathcal{M} \setminus \setminus \text{ref CALCULATE } \mathcal{M}$ 
3:    $[g_1.clients, \dots, g_n.clients] \leftarrow \text{SpectralClustering}(\mathcal{M})$ 
4:    $G \leftarrow [g_1, \dots, g_n]$ 
5:   return  $G$ 

```

with a total bandwidth of \mathcal{W} , for client c_i , its signal-to-noise ratio (SNR) is defined as follows [29]:

$$SNR_i = \frac{P_i P_L^i}{N_0 \mathcal{W}}, \quad (5)$$

where P_i is the transmission power (unit: dbm) from client c_i to its group g_x , N_0 is the thermal noise variance (unit: dbm/hz), P_L^i is the path loss (unit: db) between client c_i and group g_x , we define P_L^i as follows:

$$P_L^i = 100.7 + 23.5 \lg R, \quad (6)$$

Where R is the distance (unit: km) between client c_i and group g_x .

Based on (5) and (6), we define the transmission speed from client c_i to group g_x as follows:

$$C_i = \gamma_i \mathcal{W} \log 2(1 + SNR_i), \quad (7)$$

where $\gamma_i \mathcal{W}$ represents the bandwidth allocated to client c_i . Based on (7), we define the communication latency of local update of client c_i as:

$$t_{cm}^i = \frac{S_{model}}{C_i}, \quad (8)$$

where S_{model} is the size of the model update. Because the transmission power of the physical facility where the group g_x is located is relatively large, we ignore the communication latency from the group g_x to the client c_i .

3) *Model Update Latency*: We only cluster the classified clients once, so we take the mathematical expectation of each device's computation latency. Therefore, in the process of a model update, the model update latency of the client c_i is defined by the following formula:

$$t^i = t_{cm}^i + t_{excp}^i, \quad (9)$$

Inspired by these CFL papers [26] [27], we calculate the cosine similarity of the gradient update between clients to obtain their similarity. We define the cosine similarity between client c_i and client c_j :

$$\text{Cosine} \langle i, j \rangle = \frac{\Delta w^i \cdot \Delta w^j}{\|\Delta w^i\| \|\Delta w^j\|}, \quad (10)$$

where Δw^i is the vector of gradient update of client c_i , similarly, Δw^j is the vector of gradient update of client c_j .

Algorithm 4 Similarity Matrix Calculation

Input: (same as Algorithm 3)

Output: similarity matrix \mathcal{M}

```

1: procedure CALCULATE  $\mathcal{M}$ 
2:    $S_r \leftarrow$  set of all clients  $R$ 
3:   NormalizeTime( $S_r$ )
4:   for each client  $c_j$  in  $S_r$  parallelly do
5:      $\Delta w_0^{c_j} \leftarrow \text{PreTrainClient}(c_j, w_0)$ 
6:   for each client  $c_j$  in  $S_r$  parallelly do
7:      $\mathcal{L} \leftarrow \text{Similarity}(c_j, S_r, \beta)$ 
8:      $\mathcal{M} \leftarrow \mathcal{M} + [[\mathcal{L}]]$ 
9:   return  $\mathcal{M}$ 

10: function SIMILARITY( $c, S, \beta$ )
11:    $\mathcal{L} \leftarrow [\beta t^c]$ 
12:   for each client  $c_j$  in  $S$  parallelly do
13:      $\text{cosine} \leftarrow \text{CosineSimilarity}(\Delta w_0^c, \Delta w_0^{c_j})$ 
14:      $\mathcal{L} \leftarrow \mathcal{L} + [\text{cosine}]$ 
15:   return  $\mathcal{L}$ 

16: function PRETRAINCLIENT( $c, w$ )
17:    $\hat{w} \leftarrow w$ 
18:    $\mathcal{B} \leftarrow$  split dataset of client  $c$  into batches of size  $B$ .
19:   for each local epoch  $e$  from 1 to  $E$  do
20:     for each local batch  $b \in \mathcal{B}$  do
21:        $w \leftarrow w - \eta \nabla L(b; w)$ .
22:    $\Delta w \leftarrow w - \hat{w}$ 
23:   return Flatten( $\Delta w$ )

24: function NORMALIZETIME( $S$ )
25:    $avg \leftarrow \sum_{c_j \in S} \frac{t_{cm}^{c_j} + t_{excp}^{c_j}}{\|S\|}$ 
26:    $var \leftarrow \sum_{c_j \in S} \frac{(t_{cm}^{c_j} + t_{excp}^{c_j} - avg)^2}{\|S\|}$ 
27:   for each client  $c_j$  in  $S$  parallelly do
28:      $t^{c_j} \leftarrow \frac{t_{cm}^{c_j} + t_{excp}^{c_j} - avg}{var}$ 

```

Building Affinity Matrix

Before we build the affinity matrix, we should build the similarity matrix. In order not to make the latency with a small value lose its effect, we need to normalize these latencies. As shown in Algorithm 4, we first calculate the variance and mean value of the model update latency of the clients to be classified, expressed by var and avg respectively, and then normalize the model update latency of the clients (line 3, Line24 to line28):

$$t_n^i = \frac{t^i - avg}{var}. \quad (11)$$

Finally, given a hyper-parameter weight β , βt_n^i which represents the trade off between model update latency and gradient update on the same client c_i .

We first leverage the (10) to calculate the cosine similarities between the client c_i and the clients to be classified. We use a vector I ($I \subset \mathbb{R}^{1 \times v}$) to represent the cosine similarities of the client c_i . v is the number of clients to be classified.

Then we connect t_n^i and vector *cosine* as vector L ($L \subset \mathbb{R}^{1 \times (v+1)}$). Finally, the vector L is used as the client's data c_i for classification. Similarly, the similarity matrix M ($M \subset \mathbb{R}^{v \times (v+1)}$) is constructed by calculating the vector L of all clients to be classified. At last, Gaussian similarity is used to construct an affinity matrix between clients based on the row vector of the similarity matrix M . The formula is as follows:

$$\text{Gaussian} \langle i, j \rangle = e^{-\frac{\|L^i - L^j\|}{2\sigma^2}}, \quad (12)$$

In this paper, we use the average aggregation strategy of *FedAvg*.

To verify our CSAFL framework is reasonable, we will evaluate it based on several real federated datasets in the next section.

V. EVALUATION

In this section, we introduce the experimental results of the CSAFL framework. We evaluate our experiments based on four open federated datasets. We show the details of our experiments in Section V-A. Then, in Section V-B, we present the performance improvement of our method compared to the baselines. In order to further verify the effectiveness of our grouping strategy and update strategy, specifically, we design comparative experiments, which are introduced in Section V-C and Section V-D, respectively. For comparison, we fix the hyper-parameter time budget T in all experiments. Specifically, each communication round in the training process of each experiment has the same time budget. Our code is based on TensorFlow [30].

A. Experimental Setup

The parameter of FL setting we used in this paper is described in III, and except these, we set model delay threshold $L = 4$. We evaluated our experiments on four federal datasets, where including two image classification datasets and a synthetic dataset, an emotion analysis dataset. We use a convex multinomial logistic regression (MCLR) model for training the first three datasets, and we use a Long Short-Term Memory (LSTM) model for training the last emotion analysis task. The statistical information of datasets and models is shown in Table I. The details are as follows:

Datasets and Models. Our experiments are based on four non-IID datasets which are class-imbalanced and data-imbalanced, and we leverage the appropriate models to train these datasets, the details follow as:

- MNIST [31]: a handwritten digits data set, which is divided into 0 to 9 the 10 categories of numbers, each data is a flatted 784 dimensional (28x28) pixel image. According to the power law, we assign data to 1000 clients, and each client has two types of digits. We use the MCLR model, and the input of the model is pixel images, and the output is 10 labels of digitals.
- FEMNIST: a 62 class federated extended MNIST dataset, which is based on EMNIST [32] dataset construction, and only sampling a to j these 10 types of lowercase

characters. We divide data into 200 clients by the power law, and each client contains five types. We also use the MCLR model, and the input and output of the model are similar to MNIST.

- Synthetic: Shamir *et al.* propose a synthetic federated dataset [33]. We set (a,b)=(0.8, 0.5) and divide data into 100 clients. Similarly, for this synthetic data set, we use the MCLR model to test it.
- Sentiment140 (Sent140) [34]: a data set based on Twitter users' emoticons to express the sentiment, each user as a client, we use a two-layer logistic regression model with 256 hidden units to train emoticons based on 100 Twitter users.

Baseline. We have two baseline methods:

- a synchronous method *FedAvg* with the time budget that we call it *T-FedAvg*.
- The asynchronous method is based on T-fedAvg, which we call it *TA-FedAvg*.

Comparison methods. In order to Significantly illustrate the performance of our method, we designed six comparison experiments. The specific FL settings are similar to our method. We will elaborate in the following sections.

Metrics. We have two metrics:

- 1) Frequency of clients' idle time: We record each client's idle time in the training process of the synchronous round. In order to evaluate the effectiveness of the CSAFL framework in mitigating the straggler effect, we evaluate the frequency of clients who wait for more than 60% of the time budget.
- 2) Accuracy: CSAFL framework is based on the clustered algorithm, and its extended comparative experiments are also divided into groups for testing. The accuracy of each client is calculated based on the group model of the group to which the client belongs, so we use the weighted test accuracy to evaluate each group model for intuitively expressing the overall performance of grouped experiments.

B. Experimental Results

As shown in Table II, we present the experimental results of our framework with the baseline methods. The results show that our method is superior to the baseline method in MNIST, FEMNIST, Synthetic. Especially, CSAFL improves test accuracy by 43.7% on synthetic(0.8, 0.5) with $L = 15000$. besides, the accuracy of *TA-FedAvg* is better than that of *T-FedAvg*, which indicates the influence of the straggler effect on synchronous FL. Although the average test accuracy of CSAFL is not better than that of *T-FedAvg* on Sent140, we can observe from Fig. 4 that the group curve of CSAFL is superior.

More detail that each communication round's test accuracy is shown in Fig. 4 and Fig. 6. We observe that our method is roughly as fast as synchronous FL and faster than asynchronous FL in the convergence speed of the model on FEMNIST. That is, our CSAFL framework can converge infinite

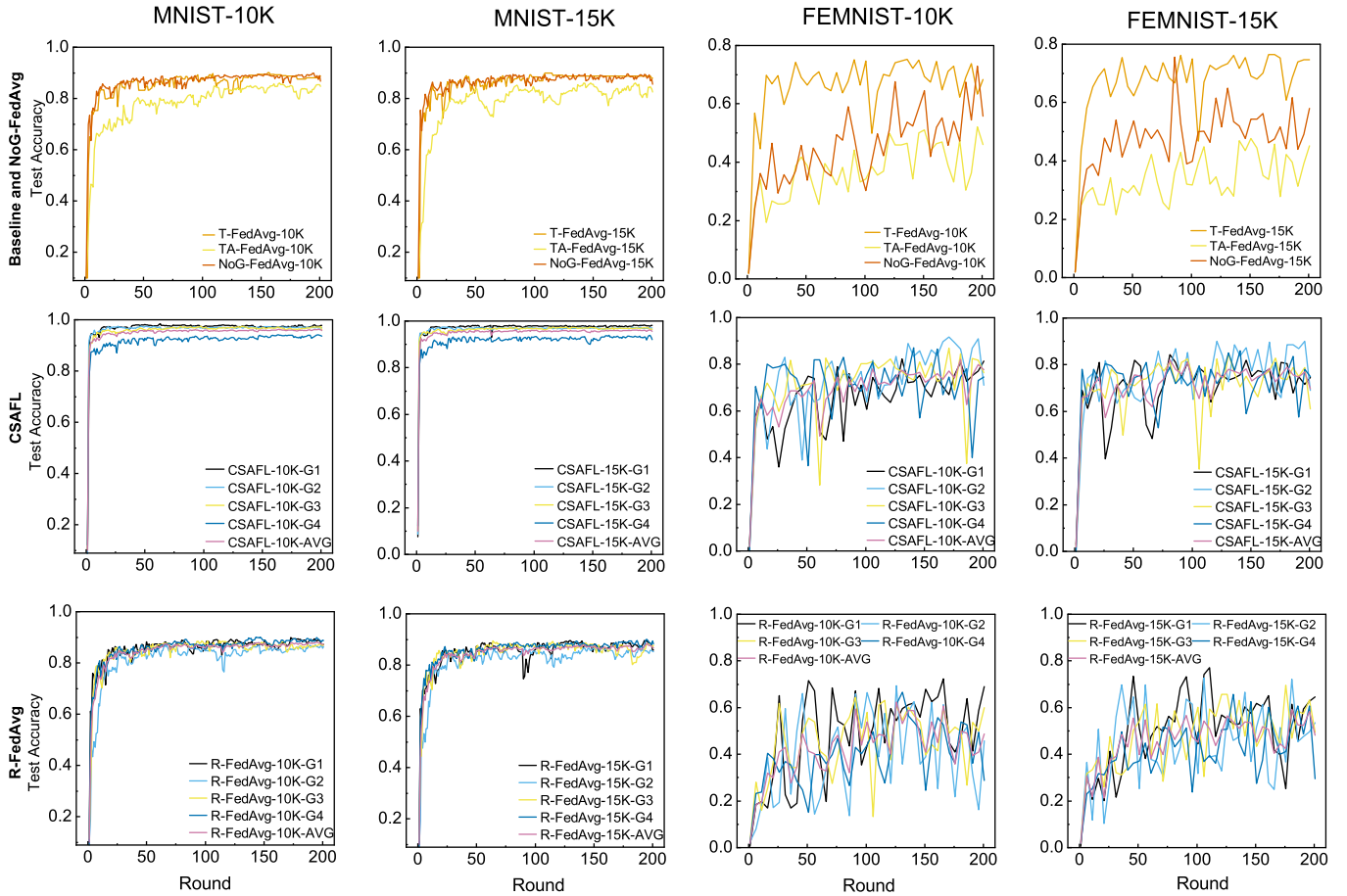


Fig. 4. The accuracy curves of CSAFL, *R-FedAvg*, baselines and *NoG-FedAvg* on MNIST and FEMNIST. 10K=10000, 15K=15000, unit:ms.

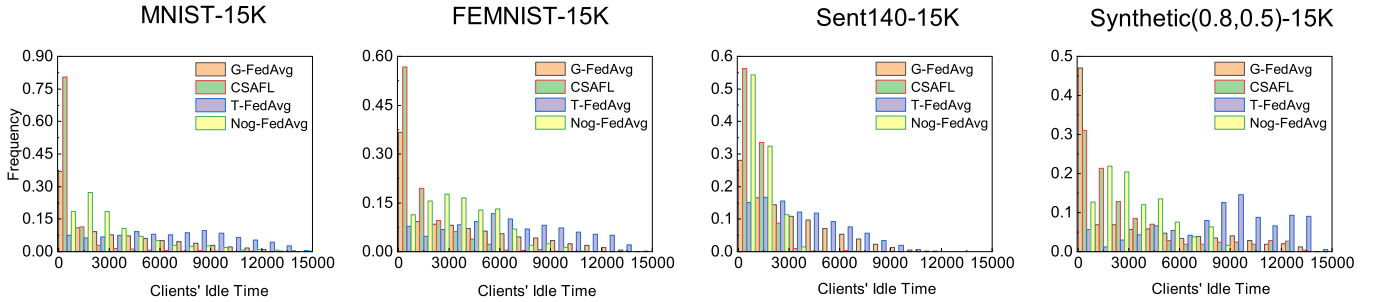


Fig. 5. The frequency distribution of the clients' idle time. 10K=10000, 15K=15000, unit: ms.

communication rounds. On the other hand, the convergence rate of CSAFL is about the same as that of the *T-FedAvg* and *TA-FedAvg* on MNIST, which may be that different model update strategies will not cause significant divergence of gradients on MNIST.

C. Effects of The Grouping Strategy

To demonstrate the effectiveness of the grouping strategy, we set up two groups of comparative experiments, as follows:

- Under the same update strategy, we use the random grouping method *R-FedAvg* based on *T-FedAvg* to compare with CSAFL. As shown in Table II, the test accuracy of CSAFL is higher than that of the *R-FedAvg* at least

7.9% on MNIST, FEMNIST, Synthetic. As for Sent140, We can observe that in Fig. 4, the curve of CSAFL is higher than *R-FedAvg* in overall accuracy. In short, our grouping strategy is effective.

- Similarly, based on the same update strategy, we design a *NoG-FedAvg* without grouping. Compared with CSAFL, the results show that the curve fluctuation of CSAFL is smaller than that of *NoG-FedAvg* in Fig. 4 and Fig. 6.

D. Effects of The Update Strategy

In order to verify the effectiveness of the update strategy, we designed two groups of comparative experiments:

TABLE II
CSAFL AND ITS COMPARISON ALGORITHMS ARE EVALUATED ON FOUR FEDERATED DATASETS, AND THIS TABLE SHOWS THE ACCURACY OF THE MODEL WITH THESE ALGORITHMS

Dataset-Time Budget	T-FedAvg	TA-FedAvg	G-FedAvg	GA-FedAvg	R-FedAvg	NoG-FedAvg	CSAFL
MNIST-10000	90.0	86.1	96.3	95.8	88.0	89.9	96.2
MNIST-15000	90.0	86.0	96.2	95.6	88.3	89.6	96.2
FEMNIST-10000	77.1	55.1	88.6	77.1	62.4	74.1	84.2
FEMNIST-15000	77.4	51.1	89.7	77.7	63.0	77.2	85.5
Synthetic-10000	26.6	33.0	88.1	64.8	31.9	40.5	66.2
Synthetic-15000	20.4	53.6	71.4	62.3	35.6	52.9	64.1
Sent140-10000	71.3	62.3	71.0	63.6	66.4	69.5	68.0
Sent140-15000	71.1	62.5	69.7	62.8	65.3	70.1	69.0

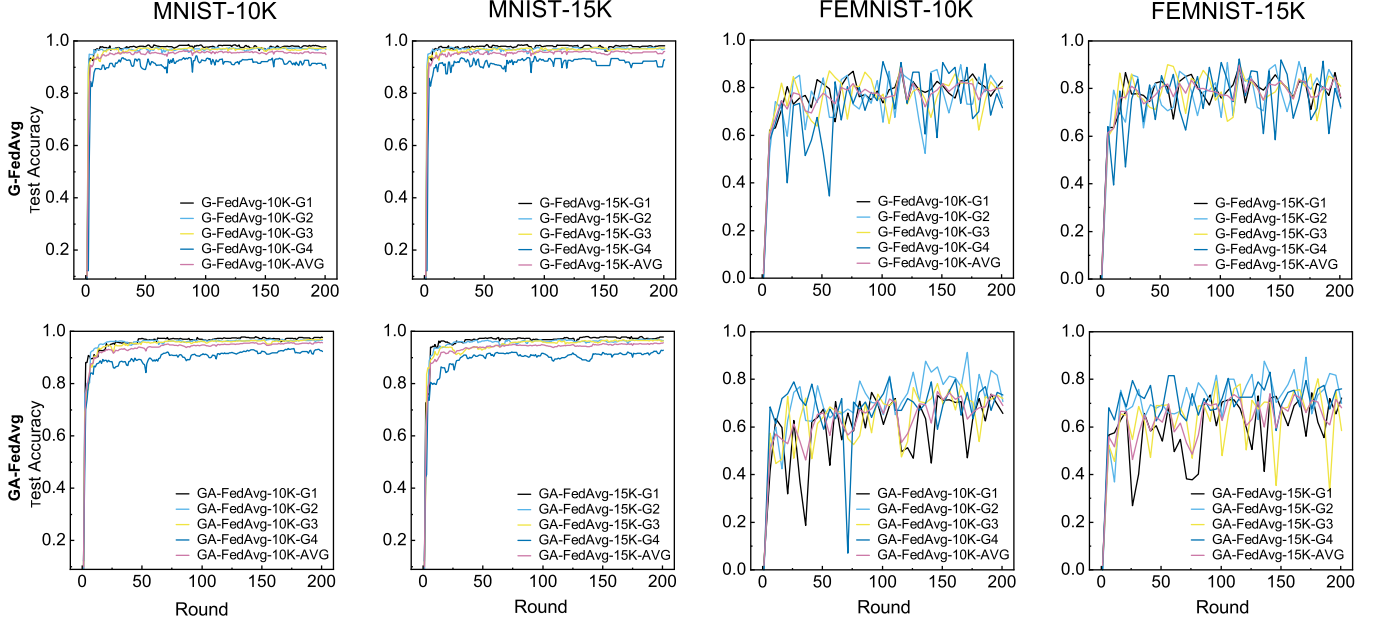


Fig. 6. The accuracy curves of *G-FedAvg*, *GA-FedAvg* on MNIST and FEMNIST. 10K=10000, 15K=15000, unit: ms.

- For different update strategies in the case of grouping, we design the synchronous FL and asynchronous FL based on FedAvg, which are named *G-FedAvg* and *GA-FedAvg*, respectively. As shown in Table II, we can observe that our method's experimental results are better than other update strategies.
- Without grouping, we use *GA-FedAvg* and *T-FedAvg* to compare with *TA-FedAvg*, and the results show the effectiveness of our strategy

E. effects of the update strategy

In order to verify that CSAFL can mitigate the straggler effect, as shown in Fig. 5, in the time budget $L = 15000ms$, we can observe the frequency distribution of clients' idle time. The results show that CSAFL can significantly mitigate the straggler effect. Especially on the MNIST, FEMNIST, Sent140, these frequencies of the clients' idle time is close to zero after $6000ms$. On the Synthetic, we can observe that the frequencies of T-FedAvg's clients' idle time are greater than 0.07 after $9000ms$, which also shows that the conclusion in Section D is correct. That is, the straggler effect affects the accuracy of the synchronization updates.

VI. CONCLUSION

In this work, we propose a new clustered semi-asynchronous federated learning framework (CSAFL), which can effectively mitigate the straggler effect and improve the accuracy of the asynchronous FL. Based on four datasets, our evaluation experiments show that the CSAFL framework is better than asynchronous FL on accuracy, convergence speed, and CSAFL is an effective solution to the straggler effect in synchronous FL. We further prove the effectiveness of the grouping strategy and update mechanism.

In the future, we will explore the following directions:

- 1) The advantages of our framework in privacy protection.
- 2) We will further expand the trade-off of time and direction; personalized learning beta is based on various data sets.
- 3) We will explore a better strategy to aggregate global models.

ACKNOWLEDGMENT

We would like to thank the reviewers for their valuable feedback and improvements to this paper. This work is partially supported by grants from the National Natural Science

Foundation of China (61601067, 61802038), Chongqing High-Tech Research Key Program (cstc2019jcsx-mbdx0063), the Fundamental Research Funds for the Central Universities under Grant (0214005207005 and 2019CDJGFJSJ001), China Postdoctoral Science Foundation (2017M620412), Chongqing Postdoctoral Special Science Foundation (XmT2018003), Chongqing Youth Talent Support Program, and Chongqing Science Fund for Distinguished Young Scholars.

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.
- [2] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
- [3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [4] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards efficient and privacy-preserving federated deep learning," in *2019 IEEE International Conference on Communications, ICC 2019, Shanghai, China, May 20-24, 2019*. IEEE, 2019, pp. 1–6.
- [5] C. Fang, Y. Guo, N. Wang, and A. Ju, "Highly efficient federated learning with strong privacy preservation in cloud computing," *Comput. Secur.*, vol. 96, p. 101889, 2020.
- [6] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, 2021.
- [7] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, "PMF: A privacy-preserving human mobility prediction framework via federated learning," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 10:1–10:21, 2020.
- [8] T. T. Vu, D. T. Ngo, H. Q. Ngo, M. N. Dao, N. H. Tran, and R. H. Middleton, "User selection approaches to mitigate the straggler effect for federated learning on cell-free massive MIMO networks," *CoRR*, vol. abs/2009.02031, 2020.
- [9] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *2020 IEEE International Conference on Communications, ICC 2020, Dublin, Ireland, June 7-11, 2020*. IEEE, 2020, pp. 1–6.
- [10] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.
- [11] M. Duan, D. Liu, X. Chen, Y. Tan, J. Ren, L. Qiao, and L. Liang, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," in *Proceedings of the IEEE 37th International Conference on Computer Design (ICCD)*. IEEE, 2019, pp. 246–254.
- [12] I. G. Cong Xie, Sanmi Koyejo, "Asynchronous federated optimization," *CoRR*, vol. abs/1903.03934, 2019. [Online]. Available: <http://arxiv.org/abs/1903.03934>
- [13] Y. Chen, Y. Ning, and H. Rangwala, "Asynchronous online federated learning for edge devices," *CoRR*, vol. abs/1911.02134, 2019.
- [14] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 13, no. 9, pp. 1088–1096, 1994.
- [15] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 219–232, 2021.
- [16] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "COTAF: convergent over-the-air federated learning," in *IEEE Global Communications Conference, GLOBECOM 2020, Virtual Event, Taiwan, December 7-11, 2020*. IEEE, 2020, pp. 1–6.
- [17] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of the 3rd SysML Conference*, 2020.
- [18] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [19] S. Dhakal, S. Prakash, Y. Yona, S. Talwar, and N. Himayat, "Coded federated learning," *CoRR*, vol. abs/2002.09574, 2020.
- [20] W. Wu, L. He, W. Lin, R. Mao, and S. A. Jarvis, "SAFA: a semi-asynchronous protocol for fast federated learning with low overhead," *CoRR*, vol. abs/1910.01355, 2019.
- [21] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 4229–4238, 2020.
- [22] X. Lu, Y. Liao, P. Liò, and P. Hui, "Privacy-preserving asynchronous federated learning mechanism for edge network computing," *IEEE Access*, vol. 8, pp. 48 970–48 981, 2020.
- [23] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.
- [24] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.
- [25] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 199–204, 1982.
- [26] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, 2020, pp. 1–9.
- [27] F. Sattler, K. Müller, T. Wiegand, and W. Samek, "On the byzantine robustness of clustered federated learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 8861–8865.
- [28] K. Lee, M. Lam, R. Pedarsani, D. S. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, 2018.
- [29] L. Liang, G. Feng, and Y. Zhang, "Integrated interference coordination for relay-aided cellular ofdma system," in *2011 IEEE International Conference on Communications (ICC)*, 2011.
- [30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, and M. Devin, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*. USENIX Association, 2016, pp. 265–283.
- [31] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, vol. 2, 2010.
- [32] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: Extending mnist to handwritten letters," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.
- [33] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1000–1008.
- [34] T. Sahni, C. Chandak, N. R. Chedeti, and M. Singh, "Efficient twitter sentiment classification using subjective distant supervision," in *9th International Conference on Communication Systems and Networks, COMSNETS 2017, Bengaluru, India, January 4-8, 2017*. IEEE, 2017, pp. 548–553.