# Supplementary Material

**Jie MA**[1], **Tianyi Zhou**[2], **Guodong Long**[1], **Jing Jiang**[1], **Chengqi Zhang**[1]
[1]Australian Artificial Intelligence Institute, FEIT, University of Technology Sydney
[2]University of Maryland
jie.ma-5@student.uts.edu.au, tianyi@umd.au
{guodong.long, jing.jiang, chengqi.zhang}@uts.edu.au

## 1 More Experimental Results

### 1.1 Ablation Study of Warmup and Extra Cost

**Impact of Warmup Rounds** As shown in Table 1 below, we gradually increase the rounds of the warmup stage (from 0 to 50) while keeping the total budget of rounds to 100 (warmup + training), considering the limited capacity of computation and communication for local devices in FL. The best performance is achieved when the warmup rounds are set to 20. However, the performance shows minimal variation when the number is set to 10, 20, 30, or 40. It demonstrates that the performance is stable when we choose warmup rounds in the area from 10 to 40. The choice of warmup round numbers exhibits low sensitivity, like on the parameter plateau.

Notably, with no warmup rounds, performance is substantially decreased due to the impact of worse-performed initial candidates of the FL system. Similarly, when the warmup rounds are increased to 50, indicating insufficient training, the performance will drop accordingly. We need to ensure there are enough training rounds with a proper number of warmup rounds.

In summary, a few warmup rounds can improve the stability of FL optimization and accuracy-related performance. Given a proper area, choosing warmup rounds is low sensitivity to performance.

**Extra Cost of integrating proposed CAM framework to existing FL methods** For simplicity, we use "FedAvg" as the measuring unit or benchmark for the cost of storage, communication and computation on local devices. In general, CAM will bring one extra 'FedAvg' cost to the existing FL methods every communication round.

As for IFCA [1], which needs to transmit $K$ cluster-specific models to each client to compute the clustering, applying our proposed CAM framework with IFCA, we need to transmit $K$ cluster models and one extra global model to the clients, that is $K + 1$ models in total. The communication cost and storage cost are listed in Table 1. Moreover, the warmup stage only incurs one "FedAvg" cost. Therefore, integrating CAM can even reduce the overall cost by increasing the number of warmup rounds.

Lastly, considering the tradeoff between performance and cost, we choose 30 warmup rounds out of 100 as the default experiment setting.

### 1.2 Experimental Results for PathMINST and TissueMNIST

Table 2 and 3 are test results for PathMNIST and TissueMNIST in cluster-wise and client-wise non-IID settings, respectively. And we have similar results as Fashion-MNIST and CIFAR-10.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Ablation study of warmup round numbers for performance and cost using "FedAvg" as the measuring unit (Other settings: CIFAR-10 dataset, IFCA, client-wise non-iid with Dirichlet distribution $\alpha = 0.1$, Cluster number $K = 10$).

| Baseline | # Warmup + Training | Performance/% | | Cost/"FedAvg" | | |
|---|---|---|---|---|---|---|
| | | Accuracy | Macro-F1 | Storage | Communication | Computation |
| IFCA | 0+100 | 47.62 | 23.36 | **10x** | 10x | 10x |
| IFCA-CAM | 0+100 | 63.75 | 32.17 | 11x | 11x | 11x |
| | 10+90 | 72.69 | 41.24 | 11x | 10x | 10x |
| | 20+80 | **73.83** | **44.72** | 11x | 9x | 9x |
| | **30+70** | 72.54 | 42.86 | 11x | 8x | 8x |
| | 40+60 | 72.98 | 42.20 | 11x | 7x | 7x |
| | 50+50 | 65.74 | 26.63 | 11x | **6x** | **6x** |

Table 2: Test results (mean±std) in **cluster**-wise non-IID settings on PathMNIST & TissueMNIST.

| Datasets | | PathMNIST | | | | TissueMNIST | | | |
|---|---|---|---|---|---|---|---|---|---|
| Non-IID setting | | $\alpha = (0.1, 10)$ | | $(3, 2)-$class | | $\alpha = (0.1, 10)$ | | $(3, 2)-$class | |
| **#Cluster** | Methods | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 |
| **1** | FedAvg | 31.38±8.58 | 14.47±4.27 | 21.36±5.48 | 11.49±2.38 | 49.96±3.39 | 18.31±4.31 | 53.46±2.21 | 15.28±1.36 |
| | FedProx | 27.6±6.15 | 14.07±3.42 | 25.7±8.48 | 11.62±1.08 | 49.78±2.64 | 17.85±3.81 | 54.92±3.7 | 15.15±1.47 |
| **5** | FedAvg+ | 35.84 | 17.01 | 25.51 | 12.14 | 49.52 | 17.59 | 54.98 | 15.12 |
| | FedProx+ | 27.57 | 15.74 | 29.7 | 13.05 | 48.88 | 17.08 | 52.24 | 15.54 |
| | IFCA | 38.13±2.53 | 25.22±1.74 | 34.16±3.76 | 22.52±1.13 | 27.44±16.39 | 16.37±10.45 | 41.87±20.04 | 21.59±7.3 |
| | IFCA-CAM | 50.12±0.42 | 25.22±3.67 | 68.45±5.83 | 39.31±1.57 | **83.08±2.16** | **39.81±3.7** | **83.26±6.47** | 36.03±0.96 |
| | FeSEM | 59.85±1.45 | 33.5±4.08 | 66.37±7.19 | 41.34±4.12 | 72.38±1.81 | 36.79±1.06 | 70.62±2.41 | 28.43±2.54 |
| | FeSEM-CAM | **70.01±1.23** | **44.09±4.94** | **71.5±2.2** | **43.69±2.96** | 80.28±4.04 | 34.77±0.49 | 75.04±4.95 | **39.33±3.32** |
| **10** | FedAvg+ | 33.19 | 19.98 | 24.82 | 13.73 | 49.5 | 18.03 | 54.78 | 13.23 |
| | FedProx+ | 28.21 | 16.17 | 35.62 | 15.95 | 46.57 | 16.47 | 53.47 | 14.88 |
| | IFCA | 42.34±2.73 | 29.1±1.52 | 37.22±4.23 | 20.2±2.04 | 38.76±10.94 | 20.38±2.01 | 49.31±13.97 | 21.51±3.68 |
| | IFCA-CAM | 66.5±3.46 | 38.12±2.32 | 66.22±3.85 | 40.75±2.2 | 81.53±7.24 | 43.88±6.98 | 88.77±15.23 | 46.48±4.98 |
| | FeSEM | 79.31±0.72 | 48.14±0.23 | 71.37±1.5 | 53.78±2.21 | 77.12±1.68 | 47.69±3.1 | 77.92±1.53 | 45.68±6.71 |
| | FeSEM-CAM | **85.06±2.62** | **61.82±5.38** | **76.41±2.22** | **64.33±4.92** | **84.91±2.83** | **53.08±1.77** | **90.22±3.03** | **60.62±2.64** |

Table 3: Test results (mean±std) in **client**-wise non-IID settings on PathMNIST & TissueMNIST.

| Datasets | | PathMNIST | | | | TissueMNIST | | | |
|---|---|---|---|---|---|---|---|---|---|
| Non-IID setting | | $\alpha = 0.1$ | | $2-$class | | $\alpha = 0.1$ | | $2-$class | |
| **#Cluster** | Methods | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 |
| **1** | FedAvg | 26.41±9.15 | 14.29±3.08 | 26.11±8.51 | 13.05±2.33 | 52.42±4.04 | 16.23±3.81 | 54.11±2.28 | 14.51±1.28 |
| | FedProx | 27.61±7.38 | 13.97±2.6 | 28.77±8.33 | 12.16±2.27 | 53.42±4.29 | 15.84±3.41 | 54.51±3.26 | 14.43±1.36 |
| **5** | FedAvg+ | 32.68 | 15.03 | 29.8 | 13.02 | 53.15 | 16.51 | 54.63 | 14.57 |
| | FedProx+ | 33.19 | 15.66 | 30.51 | 13.49 | 53.56 | 17.89 | 55.03 | 14.78 |
| | IFCA | 38.13±2.53 | 25.22±1.74 | 34.16±3.76 | 22.52±1.13 | 38.76±10.94 | 20.38±2.01 | 49.31±13.97 | 21.51±3.68 |
| | IFCA-CAM | 64.67±6.17 | 34.86±3.95 | 62.72±3.54 | 37.5±4.67 | 84.72±0.95 | 41.86±1.38 | **73.71±0.97** | **33.02±2.64** |
| | FeSEM | 59.85±1.45 | 33.5±4.08 | 64.46±6.12 | 38.41±3.19 | 72.88±1.11 | 33.19±1.7 | 70.62±2.41 | 28.43±2.54 |
| | FeSEM-CAM | **68.81±1.29** | **49.22±2.2** | **68.92±1.13** | **47.32±1.99** | **87.88±1.27** | **45.82±1.54** | 70.09±0.86 | 29.49±0.77 |
| **10** | FedAvg+ | 29.83 | 16.75 | 28.35 | 13.49 | 53.5 | 18.03 | 54.58 | 13.46 |
| | FedProx+ | 29.36 | 16.55 | 29.07 | 13.63 | 54.69 | 17.36 | 56.03 | 15.21 |
| | IFCA | 51.88±13.67 | 27.81±2.21 | 37.22±4.23 | 20.2±2.04 | 27.44±16.39 | 16.37±10.45 | 41.87±20.04 | 21.59±7.3 |
| | IFCA-CAM | 77.32±1.0 | 54.89±3.42 | 67.91±3.21 | 40.49±3.58 | **88.24±1.62** | 54.12±4.15 | 74.5±0.89 | 32.04±1.17 |
| | FeSEM | 78.93±4.27 | 52.94±5.42 | 70.93±4.27 | 52.94±5.42 | 78.85±2.29 | 52.32±7.59 | 77.92±1.53 | 45.68±6.71 |
| | FeSEM-CAM | **82.38±2.6** | **63.84±2.03** | **72.95±0.36** | **54.44±1.05** | 87.09±1.97 | **54.77±2.2** | **80.13±1.6** | **51.9±2.15** |

### 1.3 Comparison with Ensemble and Finetune

In Table 4, we further analyze CAM under various scenarios. The terms "-Finetune" and "+" denote finetuning base methods for one additional round and ensembling both methods via soft voting, respectively. We present a few examples as follows.

- **FedAvg+IFCA:** Initially, we separately train FedAvg and IFCA on the same partitioned dataset for 100 rounds, keeping all other hyperparameters identical. We then ensemble the trained models of FedAvg and IFCA to test on the relevant clients using soft voting. The inference is carried out using the formula below, which aligns with the inference method in Fed-CAM,

$$\operatorname{argmax} y = f(x; \Theta_g) + f(x; \Theta_{c(i)}). \tag{1}$$

- **FedAvg-Finetune+IFCA-Finetune:** Similar to the previous method, we train FedAvg and IFCA separately on the same partitioned dataset for 100 rounds, and then finetune each

locally for one additional round. Next, we ensemble the trained models of FedAvg-Finetune and IFCA-Finetune to test on the relevant clients using soft voting.

- **IFCA-CAM-Finetune:** After obtaining $\Theta_g$ and $\Theta_{c(i)}$, we finetune both locally for one round without aggregation. Then, we use the finetuned models for testing, applying the same inference method as before.

Table 4: More comparison, CIFAR-10 cluster-wise non-IID (Dirichlet), $K = 10$

| Methods | Accuracy | Macro-F1 |
|---|---|---|
| FedAvg | 24.38±3.30 | 11.69±3.15 |
| IFCA | 34.84±5.82 | 22.76±3.99 |
| FedAvg+IFCA | 35.62±4.73 | 24.31±3.65 |
| IFCA-CAM | 70.9±1.18 | 40.03±1.28 |
| FedAvg-Finetune+IFCA-Finetune | 65.89± 2.31 | 39.51±1.94 |
| IFCA-CAM-Finetune | **78.97± 1.64** | **52.3± 2.42** |
| FeSEM | 66.89±2.18 | 38.35±4.24 |
| FedAvg+FeSEM | 67.37±1.85 | 42.03 ±2.45 |
| FeSEM-CAM | 78.45±1.71 | 49.5±1.13 |
| FedAvg-Finetune+FeSEM-Finetune | 77.63±1.84 | 50.34±2.58 |
| FeSEM-CAM-Finetune | **81.33± 1.51** | **57.64± 2.17** |

### 1.4 More Clustering Analysis

**Clustering stability** Figure 1 demonstrates that the clustering results remain stable after five communication rounds.

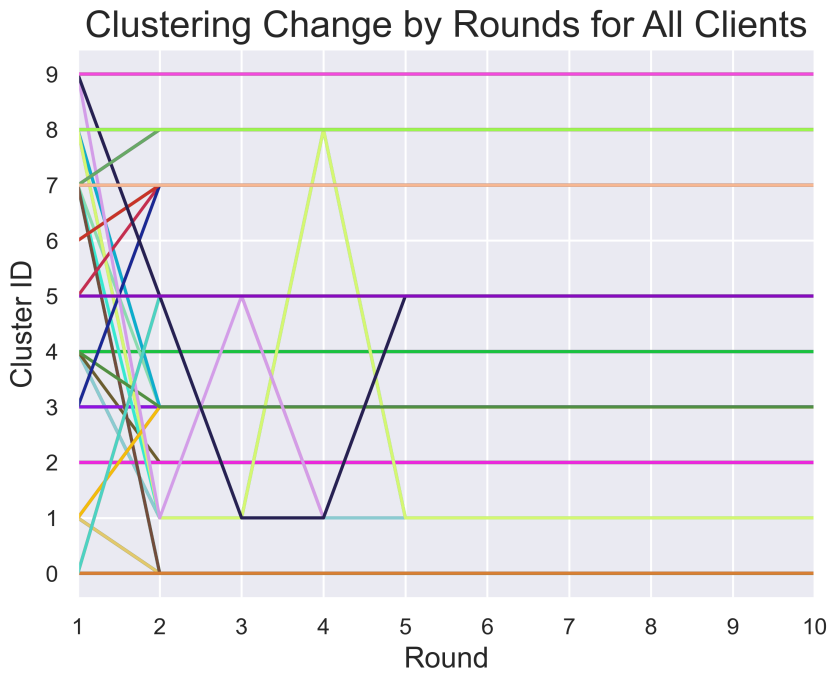


Figure 1: A Clustering change example for IFCA-CAM with client-wise non-IID and $K = 10$ on CIFAR-10. Note that there are 200 lines in this graph, and each represents a client. The bold line in this figure is the combination of lines of clients within one cluster. **After five rounds, the clustering remains stable.**

**Clustering accuracy in highly-skewed cluster-wise non-iid setting** Figure 2 is an example of highly-skewed cluster-wise non-iid setting with cluster size $\{10, 10, 10, 10, 10, 20, 30, 30, 70\}$. Then

Figure 3 shows the difference between clustering results when stable and ground truth. Compared with clustering collapse in IFCA, in which all clients fall into one cluster, IFCA-CAM can reveal most of the clustering ground truth.

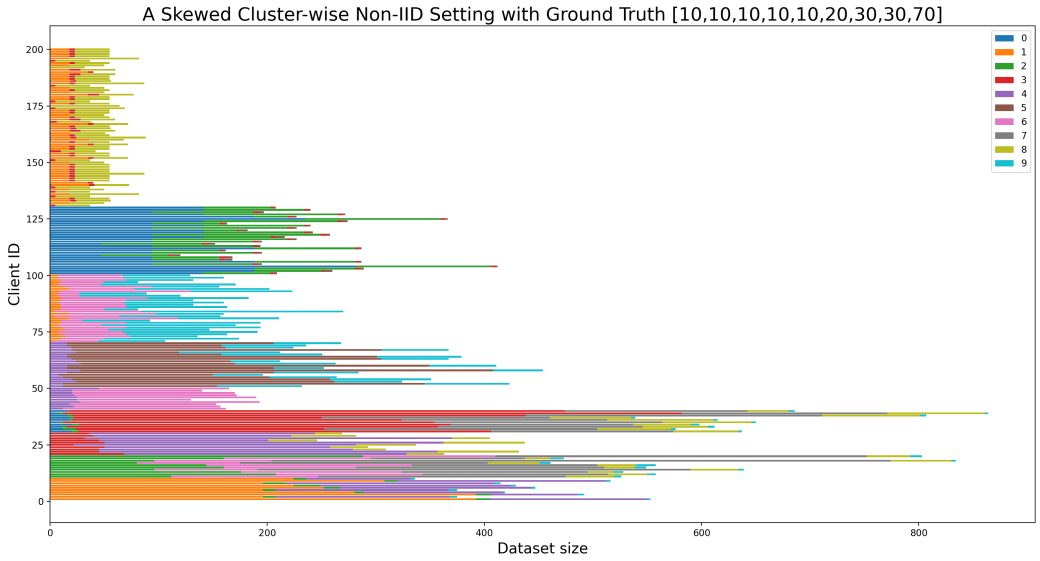


Figure 2: A skewed non-iid setting example on CIFAR-10. Legends represent labels of the dataset.

## 2 Convergence Proof

**Claim 1.** *(Identical data distribution with one cluster for FedSEM-CAM). Assume that clients clustered into the same cluster have the same data distribution when clustering is stable, especially in FedSEM-CAM.*

*Remark* 1. Claim 1 can be validated by experimental analysis of clustering in this paper easily, as FeSEM-CAM uses the parameters of the last layers for clustering, which contains label distribution information of clients.

**Lemma 1.** *(Bounding $\epsilon_{distriubte}$). Under the assumption of convexity of cluster models, and the claim of dentical data distribution with one cluster for FedSEM-CAM, we can get*

$$\epsilon_{distriubte} = \mathcal{L}(\Theta_g, \Theta_{c(i)}) - \mathcal{L}(\Theta_g, \theta_i, c(i)) \tag{2}$$

$$\leq 0. \tag{3}$$

*Proof.* For minloss-based methods, it is straightforward to prove that $\epsilon_{distriubte} \leq 0$. However, for distance-based methods like FeSEM-CAM, bounding $\epsilon_{distriubte}$ may require the introduction of a new bound in Lemma 4 of work [2]. According to the assumptions of convexity and identical distribution within one cluster, we have

$$\mathbb{E}_t[\mathcal{L}(\Theta_g, \Theta_{c(i)}) - \mathcal{L}(\Theta_g, \theta_i)] \tag{4}$$

$$= \sum_{k=1}^{K} \sum_{c(i)=k} \frac{n_i}{n} \mathbb{E}_t[\ell(\Theta_g, \Theta_{c(i)}) - \ell(\Theta_g, \theta_i)] \tag{5}$$

$$= \sum_{k=1}^{K} \frac{n_k}{n} \sum_{c(i)=k} \frac{n_i}{n_k} \mathbb{E}_t[\ell(\Theta_g, \sum_{i \in C_k} \frac{n_i}{n_k} \theta_i) - \ell(\Theta_g, \theta_i)] \tag{6}$$

$$\leq 0, \tag{7}$$

where $n_k$ is the number of clients in Cluster $k$, and $\sum_{k=1}^{K} n_k = n$. □

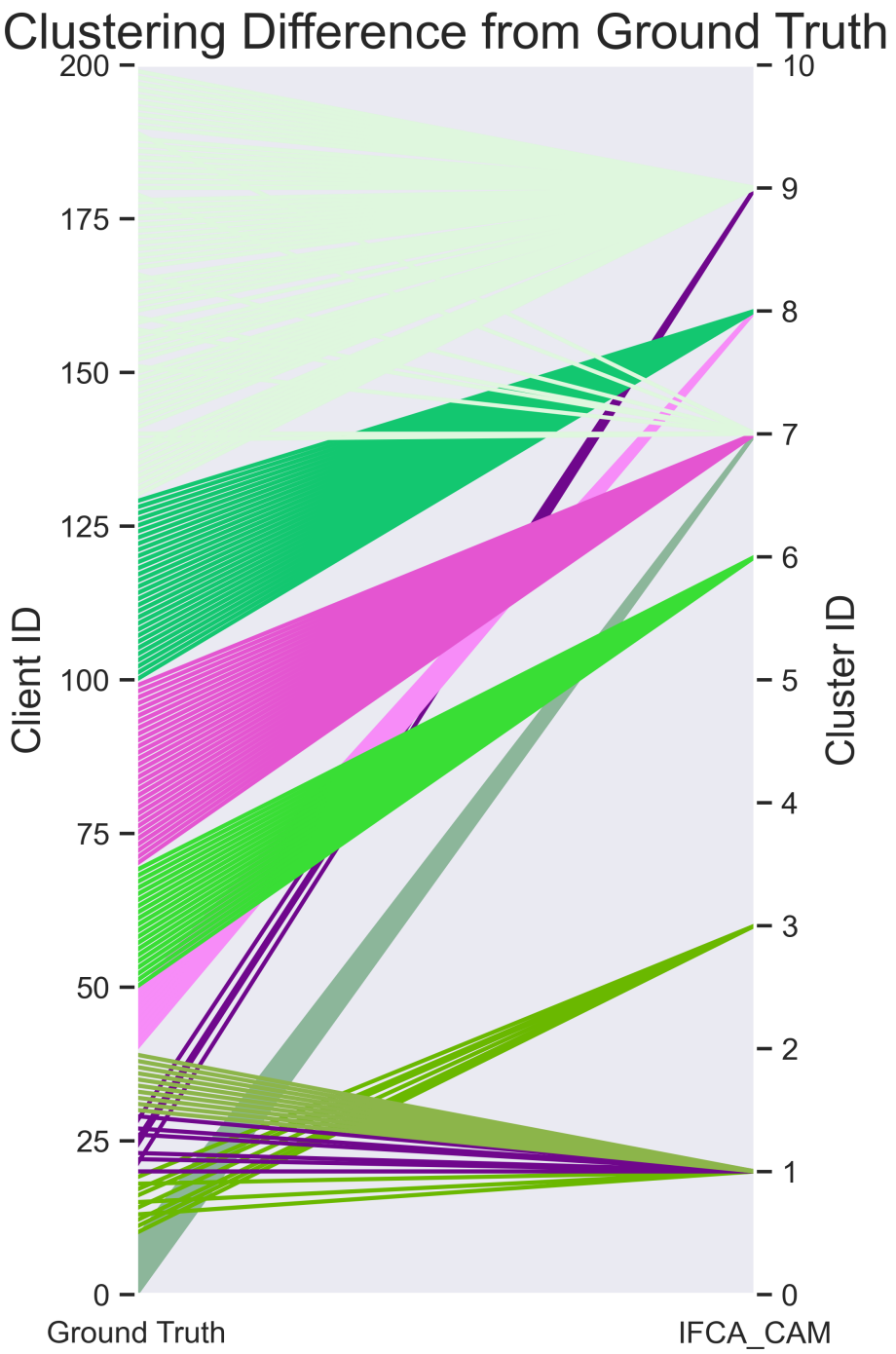


Figure 3: In the context of the highly-skewed clustering scenario depicted in Figure 2, the differences between IFCA-CAM's clustering and the actual ground truth remain minimal. Conversely, the clustering of IFCA easily collapses into a single cluster. The right y-axis indicates the cluster id. The color represents the ground truth, while the lines indicate the transition from the original ground truth to the clustering through CAM. Notably, **CAM also demonstrates its capability to alleviate clustering collapse and imbalance in skewed clustering settings successfully.**

Finally, the convergence proof is as follows.

*Proof.* Firstly, we simplify the objective function to minimize as below,

$$\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m} \ell(\Theta_g, \Theta_{c(i)}). \tag{8}$$

Then the proof outline is as follows,

$$\mathcal{L}(\Theta_g^{(t+1)}, \Theta_{c'(i)}^{(t+1)}) - \mathcal{L}(\Theta_g^{(t)}, \Theta_{c(i)}^{(t)}) \tag{9}$$

$$= \underbrace{\mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c(i)) - \mathcal{L}(\Theta_g^{(t)}, \Theta_{c(i)}^{(t)})}_{\epsilon_{fedsim}} + \underbrace{\mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c'(i)) - \mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c(i))}_{\epsilon_{cluster}} \tag{10}$$

$$+ \underbrace{\mathcal{L}(\Theta_g^{(t+1)}, \Theta_{c'(i)}^{(t+1)}) - \mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c'(i))}_{\epsilon_{distribute}}, \tag{11}$$

where $c'(i)$ represents the assign relationships of round $t+1$ compared to $c(i)$ of round $t$ and

$$\Theta_{c'(i)}^{(t+1)} \leftarrow \sum_{i \in C_k} \frac{n_i}{\sum_{i \in C_k} n_i} \theta_i^{(t+1)}, \ \forall\, c'(i) = k. \tag{12}$$

**Bounding $\epsilon_{fedsim}$.** In this process, $c(i)$ does not change. Fed-CAM can be seen doing parameter sharing for one global and parameter personalization for clients in clusters. So this process is equal to FedSim [3], then we have,

$$\mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}) - \mathcal{L}(\Theta_g^{(t)}, \Theta_{c(i)}^{(t)}) \tag{13}$$

$$\leq \underbrace{\langle \nabla_{\Theta_g} \mathcal{L}(\Theta_g^{(t)}, \Theta_{c(i)}^{(t)}), \Theta_g^{(t+1)} - \Theta_g^{(t)} \rangle}_{\epsilon_{1,g}} + \underbrace{\sum_{i=1}^{m} \langle \nabla_{\Theta_k} \ell(\Theta_g^{(t)}, \Theta_{c(i)}^{(t)}), \theta_i^{(t+1)} - \Theta_{c(i)}^{(t)} \rangle}_{\epsilon_{1,i}} \tag{14}$$

$$+ \underbrace{L\|\Theta_g^{(t+1)} - \Theta_g^{(t)}\|^2}_{\epsilon_{2,g}} + \underbrace{\sum_{i=1}^{m} L\|\theta_i^{(t+1)} - \Theta_{c(i)}^{(t)}\|^2}_{\epsilon_{2,i}}. \tag{15}$$

By mapping $\epsilon_{1,g}, \epsilon_{2,g}, \epsilon_{1,i}, \epsilon_{2,i}$ to $\tau_{1,u}, \tau_{2,u}, \tau_{1,v}, \tau_{1,v}$ respectively in the convergence proof for FedSim, with Claim 14, 15, 16, 17 in [3], we will obtain the same bound.

**Bounding $\epsilon_{cluster}$.** In this bounding step, we assign a new cluster for all clients, but distribute the cluster model later. Therefore $c(i)$ changes to $c'(i)$, but $\theta_i$ keeps the same. And we got

$$\mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c'(i)) - \mathcal{L}(\Theta_g^{(t+1)}, \theta_i^{(t+1)}, c(i)) = 0 \tag{16}$$

**Bounding $\epsilon_{distriubte}$.** According to Lemma 1, we have $\epsilon_{distriubte} \leq 0$.

Finally, combining $\epsilon_{fedsim},\ \epsilon_{cluster},\ \epsilon_{distribute}$, taking full expectation and telescoping over $t = 1, \ldots, T$, we have the same error bound and convergence rate with FedSim. □

## References


[1] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[2] Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv preprint arXiv:2202.06187*, 2022.

[3] Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pages 17716–17758. PMLR, 2022.