A. PROOF OF THEOREM

A.1. Proof of Theorem 1

In this section, we give the proofs in detail. Due to the smoothness in Assumption 1, taking the expectation of $f(w_{t+1})$ over the randomness in round t, we have

$$\mathbb{E}_t[f(w_{t+1})] \tag{12}$$

$$\leq f(w_t) + \langle \nabla f(w_t), \mathbb{E}_t[w_{t+1} - w_t] \rangle + \frac{L}{2}\mathbb{E}_t[\|w_{t+1} - w_t\|^2] \tag{13}$$

$$= f(w_t) + \langle \nabla f(w_t), \mathbb{E}_t[\eta\eta_L\Delta_t + \eta\eta_L E\nabla f(w_t) - \eta\eta_L E\nabla f(w_t)] \rangle + \frac{L}{2}\eta^2\eta_L^2\mathbb{E}_t[\|\Delta_t\|^2] \tag{14}$$

$$= f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 + \eta \underbrace{\langle \nabla f(w_t), \mathbb{E}[\eta_L\Delta_t + \eta_L E\nabla f(w_t)] \rangle}_{A_1} + \frac{L}{2}\eta^2\eta_L^2 \underbrace{\mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} \tag{15}$$

Note that the term A_1 can be bounded as follows:

$$A_1 = \langle \nabla f(w_t), \mathbb{E}_t[\eta_L\Delta_t + \eta_L E\nabla f(w_t)] \rangle \tag{16}$$

$$= \langle \nabla f(w_t), \mathbb{E}_t[\eta_L\bar{\Delta}_t + \eta_L e_t + \eta_L E\nabla f(w_t)] \rangle \tag{17}$$

$$= \left\langle \nabla f(w_t), \mathbb{E}_t\left[-\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\eta_L\nabla F^k(w_{t,\tau}^k) + \eta_L e_t + \eta_L E\frac{1}{K}\sum_{k=1}^{K}\nabla F^k(w_t) \right] \right\rangle \tag{18}$$

$$= \left\langle \sqrt{\eta_L E}\nabla f(w_t), -\frac{\sqrt{\eta_L}}{K\sqrt{E}}\mathbb{E}_t\left[\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - Ke_t \right] \right\rangle \tag{19}$$

$$\overset{(a_1)}{=} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - Ke_t \right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 \tag{20}$$

$$\overset{(a_2)}{\leq} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) \right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 + \frac{\eta_L\mathbb{E}_t\|e_t\|^2}{E} \tag{21}$$

$$\overset{(a_3)}{\leq} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\mathbb{E}_t\left\|\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)\right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 + \frac{\eta_L\mathbb{E}_t\|e_t\|^2}{E} \tag{22}$$

$$\overset{(a_4)}{\leq} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L L^2}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\mathbb{E}_t\left\|w_{t,\tau}^k - w_t\right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 + \frac{\eta_L\mathbb{E}_t\|e_t\|^2}{E} \tag{23}$$

$$\overset{(a_5)}{\leq} \eta_L E(\frac{1}{2} + 30\eta_L^2E^2L^2)\|\nabla f(w_t)\|^2 + 5\eta_L^3E^2L^2(\rho_L^2 + 6E\rho_G^2) - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\| \sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 + \frac{\eta_L\mathbb{E}_t\|e_t\|^2}{E} \tag{24}$$

where (a_1) follows from that $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2}[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2]$, (a_2) is due to that $\mathbb{E}\|x_1 + x_2\|^2 \leq 2\mathbb{E}[\|x_1\|^2 + \|x_2\|^2]$, (a_3) is due to that $\mathbb{E}\|x_1 + ... + x_n\|^2 \leq n\mathbb{E}[\|x_1\|^2 + ...\|x_n\|^2]$, (a_4) is due to Assumption 1 and (a_5) follows from Lemma 1.

The term A_2 can be bounded as

$$
\begin{aligned}
A_2 =& \mathbb{E}_t[\|\Delta_t\|^2] = \mathbb{E}_t[\|\bar{\Delta}_t + e_t\|^2] && (25)\\
\overset{(a_6)}{\leq}& 2\mathbb{E}_t\|\bar{\Delta}_t\|^2 + 2\mathbb{E}_t\|e_t\|^2 \\
\leq& \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} g_{t,\tau}^k\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 && (26)\\
\overset{(a_7)}{\leq}& \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} (g_{t,\tau}^k - \nabla F^k(w_{t,\tau}^k))\right\|^2\right] + \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k)\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 && (27)\\
\overset{(a_8)}{\leq}& \frac{2E}{K}\rho_L^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + \frac{4}{K^2}\mathbb{E}_t\left\|Ke_t\right\|^2 + 2\mathbb{E}_t\|e_t\|^2 \\
=& \frac{2E}{K}\rho_L^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + 6\mathbb{E}_t\|e_t\|^2 && (28)
\end{aligned}
$$

where both (a_6) is due to that $\mathbb{E}\|x_1 + x_2\|^2 \leq 2\mathbb{E}[\|x_1\|^2 + \|x_2\|^2]$, (a_7) follows the fact that $\mathbb{E}[\|\mathbf{x}\|^2] = \mathbb{E}[\|\mathbf{x} - \mathbb{E}\mathbf{x}\|^2] + \|\mathbb{E}\mathbf{x}\|^2$, and (a_8) is due to Assumption 3. Substituting the inequalities of A_1 and A_2 into the original inequality, we have:

$$
\begin{aligned}
& \mathbb{E}_t[f(w_{t+1})] && (29)\\
\leq& f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 + \eta \underbrace{\langle \nabla f(w_t), \mathbb{E}[\eta_L \Delta_t + \eta_L E \nabla f(w_t)]\rangle}_{A_1} + \frac{L}{2}\eta^2\eta_L^2 \underbrace{\mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} && (30)\\
\leq& f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 \\
& + \eta\eta_L E(\frac{1}{2} + 30\eta_L^2 E^2 L^2)\|\nabla f(w_t)\|^2 + 5\eta\eta_L^3 E^2 L^2(\rho_L^2 + 6E\rho_G^2) \\
& - \frac{\eta\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta\eta_L \mathbb{E}_t\|e_t\|^2}{E} \\
& + \frac{EL\eta^2\eta_L^2}{K}\rho_L^2 + \frac{2L\eta^2\eta_L^2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + 3\eta^2\eta_L^2 L\mathbb{E}_t\|e_t\|^2 && (31)\\
=& f(w_t) - \eta\eta_L E(\frac{1}{2} - 30\eta_L^2 E^2 L^2)\|\nabla f(w_t)\|^2 \\
& + 5\eta\eta_L^3 E^2 L^2(\rho_L^2 + 6E\rho_G^2) + \frac{EL\eta^2\eta_L^2}{K}\rho_L^2 + \Big(\frac{\eta\eta_L}{E} + 3\eta^2\eta_L^2 L\Big)\mathbb{E}_t\|e_t\|^2 \\
& - \left(\frac{\eta\eta_L}{2EK^2} - \frac{2L\eta^2\eta_L^2}{K^2}\right)\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 && (32)\\
\overset{(a_9)}{\leq}& f(w_t) - c\eta\eta_L E\|\nabla f(w_t)\|^2 + 5\eta\eta_L^3 E^2 L^2(\rho_L^2 + 6E\rho_G^2) + \frac{EL\eta^2\eta_L^2}{K}\rho_L^2 + \Big(\frac{\eta\eta_L}{E} + 3\eta^2\eta_L^2 L\Big)\mathbb{E}_t\|e_t\|^2 && (33)
\end{aligned}
$$

where (a_9) follows from $\left(\frac{\eta\eta_L}{2EK^2} - \frac{2L\eta^2\eta_L^2}{K^2}\right) < 0$ if $\eta\eta_L \leq \frac{1}{4EL}$, and that there exits a constant $c > 0$ satisfying $(\frac{1}{2} - 30\eta_L^2 E^2 L^2) > c > 0$ if $\eta_L < \frac{1}{\sqrt{60}EL}$.

Rearranging and summing from $t = 0, ..., T-1$, we have:

$$\sum_{t=0}^{T-1} c\eta\eta_L E\mathbb{E}\|\nabla f(w_t)\|^2 \tag{34}$$

$$\leq f(w_0) - f(w_T) + TE\eta\eta_L \left[5\eta_L^2 EL^2(\rho_L^2 + 6E\rho_G^2) + \frac{\eta\eta_L L}{K}\rho_L^2\right] + \left(\frac{\eta\eta_L}{E} + 3\eta^2\eta_L^2 L\right)\sum_{t=0}^{T=1}\mathbb{E}_t\|e_t\|^2 \tag{35}$$

which implies,

$$\min_{t=0,...,T-1}\mathbb{E}\|\nabla f(w_t)\|^2 \leq \frac{f_0 - f_*}{c\eta\eta_L ET} + \Phi + \Psi(e_0, ..., e_{T-1}) \tag{36}$$

where

$$\Phi = \frac{1}{c}\left[5\eta_L^2 EL^2(\rho_L^2 + 6E\rho_G^2) + \frac{\eta\eta_L L}{K}\rho_L^2\right] \tag{37}$$

$$\Psi(e_0, ..., e_{T-1}) = \frac{1 + 3\eta\eta_L LE}{cE^2T}\sum_{t=0}^{T-1}\mathbb{E}_t\|e_t\|^2 \tag{38}$$

This completes the proof.

A.2. Bounds on $\mathbb{E}\|e_t\|^2$

The error bound with client dropout:

$$\mathbb{E}[\|e_t\|^2] = \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k\in\mathcal{K}\setminus\mathcal{S}_t}(\tilde{\Delta}_t^k - \Delta_t^k)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k\in\mathcal{K}\setminus\mathcal{S}_t}\frac{1}{S_t}\sum_{k'\in\mathcal{S}_t}(\Delta_t^{k'} - \Delta_t^k)\right\|^2\right] \tag{39}$$

$$\leq \frac{(K - S_t)^2}{K^2}\sigma_P^2 \leq \alpha^2\sigma_P^2 \tag{40}$$

The error bound with friend model substitution (full information) :

$$\mathbb{E}[\|e_t\|^2] = \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k\in\mathcal{K}\setminus\mathcal{S}_t}(\tilde{\Delta}_t^k - \Delta_t^k)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k\in\mathcal{K}\setminus\mathcal{S}_t}(\Delta_t^{\phi_t(k)} - \Delta_t^k)\right\|^2\right] \tag{41}$$

$$\leq \frac{(K - S_t)^2}{K^2}\sigma_F^2 \leq \alpha^2\sigma_F^2 \tag{42}$$

where $\phi_t(k)$ is a friend of k that does not dropout in round t.

B. EXPERIMENT SETTINGS

We use Python3 and the Pytorch library, and our code is adapted from [20], which is under the MIT License. The experiments were run on an Ubuntu 18.04 machine with an Intel Core i7-10700KF 3.8GHz CPU and GeForce RTX 3070 GPU. All experiment results are averaged over 10 repeats.

We perform experiments on two standard public datasets, namely MNIST and CIFAR-10, which are widely used in FL experiments, in a clustered setting as well as a general setting. In the clustered settings (one on MNIST and one on CIFAR-10), we artificially create 5 client clusters where clients in the same cluster possess data samples with the same labels. Thus, clients in the same cluster are naturally regarded as friends. However, the clustering structure is *unknown* to our algorithm. Such a clustering setting provides a controlled environment for us to evaluate the friend discovery performance of FL-FDMS. In the general setting (on CIFAR-10), 20 clients receive a random subset of the whole dataset using a common way of generating non-iid FL datesets that is widely used in existing works.

B.1. FL Dataset

Clustered Setting - MNIST: The MNIST dataset has 60000 training data samples with 10 classes. The training dataset is first split into 10 sub-datasets with samples in the same sub-dataset having the same label. There are 20 clients which are grouped into 5 client clusters with an equal number of clients. Each client cluster is associated with 2 randomly drawn sub-datasets. Then each client randomly draws 200 samples from its corresponding two sub-datasets. This approach to creating the FL dataset was introduced in a recent clustered FL work.

Clustered Setting - CIFAR-10: The CIFAR-10 dataset has 50000 training data samples with 10 classes. The training dataset is first split into 10 sub-datasets with samples in the same sub-dataset having the same label. There are 20 clients which are grouped into 5 client clusters with an equal number of clients. Each client cluster is associated with 2 randomly drawn sub-datasets. Then each client randomly draws 1000 samples from its corresponding two sub-datasets.

General Setting - CIFAR-10: The CIFAR-10 dataset has 50000 training data samples. After shuffling the samples in label order, all samples are divided into 200 partitions with each partition having 250 samples. There are 20 clients. Each client then randomly picks 2 partitions. This method is a common way of generating non-i.i.d. FL dataset, which is widely used in the existing works.

B.2. FL Models

MNIST: The CNN model has two 5×5 convolution layers, a fully connected layer with 320 units and ReLU activation, and a final output layer with softmax. The first convolution layer has 10 channels while the second one has 20 channels. Both layers are followed by 2×2 max pooling. The following parameters are used for training: the local batch size $BS = 5$, the number of local iterations $E = 2$, the local learning rate $\eta_L = 0.1$ and the global learning rate $\eta = 1$.

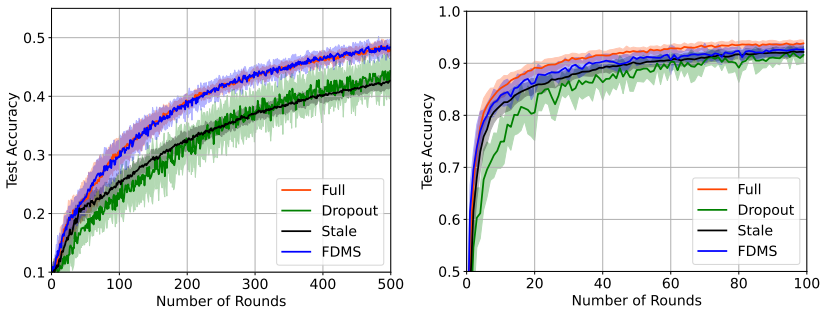
CIFAR-10: The CNN model has two 5×5 convolution layers, three fully connected layers and ReLU activation, and a final output layer with softmax. The following parameters are used for training: the local batch size $BS = 20$, the number of local iterations $E = 2$, the local learning rate $\eta_L = 0.1$ and the global learning rate $\eta = 1$.

C. ADDITIONAL EXPERIMENTS

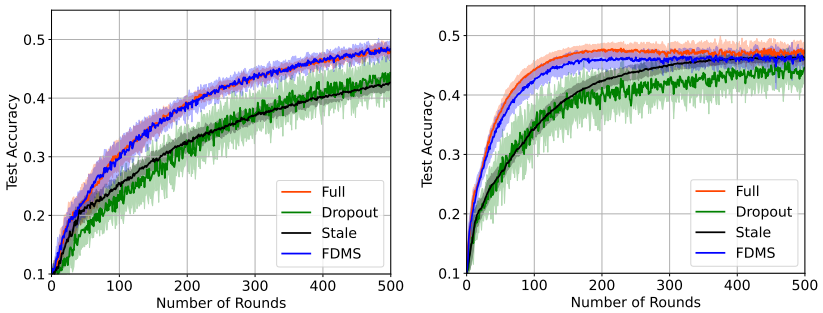
The error bound of FL-FDMS $\mathbb{E}\|e_t\|^2$ in Eq.(49) is influenced by the number of local iterations E and the number of clients K. Next, we perform additional experiments to explore their impacts.

C.1. Impact of number of local iterations E

We present more results on the performance comparison in the MNIST clustered setting and the CIFAR-10 clustered setting with different E. We fix $\alpha = 0.5$ and $K = 20$ for all the following experiments. To investigate the impact of E, we consider two values $E = 1$ and $E = 5$.



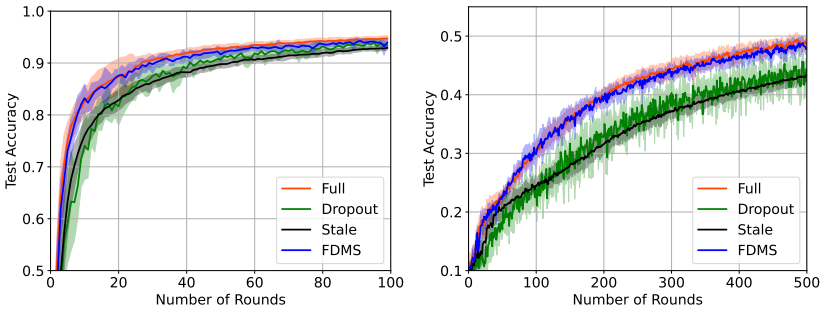
(a) MNIST ($E = 1$) (b) MNIST ($E = 5$)



(c) CIFAR-10 ($E = 1$) (d) CIFAR-10 ($E = 5$)

Fig. 4: Performance comparison with $\alpha = 0.5$ and $K = 20$

In Fig.4, we find that the **FL-FDMS** still shows superior performance in terms of test accuracy and convergence speed. However, **Dropout** and **Stale** show different trends for different E. For a larger E, using staled models tends to help the dropout situation better.



(a) MNIST ($K = 40$) (b) CIFAR-10 ($K = 40$)

Fig. 5: Performance comparison with $\alpha = 0.5$ and $E = 2$

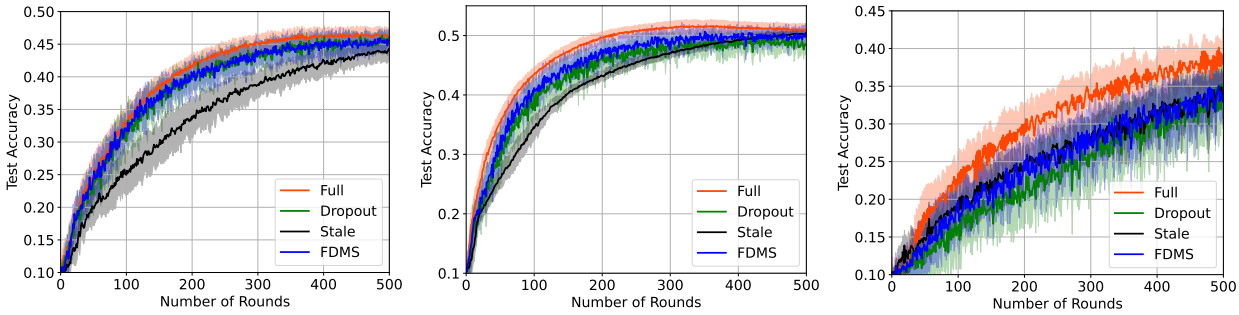
C.2. Impact of number of clients K

To investigate the impact of K, we fix $E = 2$ and increase the number of clients to $K = 40$. To keep the same total amount of data in the system, we adjust just the number of data samples on each client. For MNIST, each client now has 100 samples. For CIFAR-10, each client has 500 samples. Other settings are as described in Appendix B.

By comparing Fig.5 and the corresponding parts in Fig.1, we find that as K increases, the **FL-FDMS** outperforms **Dropout** and **Stale** even more. This is because as K increases, more clients dropout. If the model updates from dropout clients are not compensated, the global model can gradually deviate from the optimal value and eventually degrade the learning performance and affect the system stability. The additional experiments further verify that **FL-FDMS** can handle well the client dropout in FL.

C.3. Impact of non-i.i.d. level

We conducted experiments comparing different degrees of non-i.i.d.-ness, and the results are based on a general setting which is introduced in B.1. To create non-i.i.d. data with varying degrees, we divided the dataset into 100, 200, and 500 partitions for high, medium, and low degrees of non-i.i.d.-ness, respectively. Then, we randomly selected one partition for the high degree, two partitions for the medium degree, and five partitions for the low degree. In all cases, each client's dataset comprises 500 samples.



(a) CIFAR-10 (P:500, S:100, Low) (b) CIFAR-10 (P:200, S:250, Medium) (c) CIFAR-10 (P:100, S:500, High)

Fig. 6: Performance comparison on the CIFAR-10 general setting ($\alpha = 0.5$) with various non-i.i.d. level

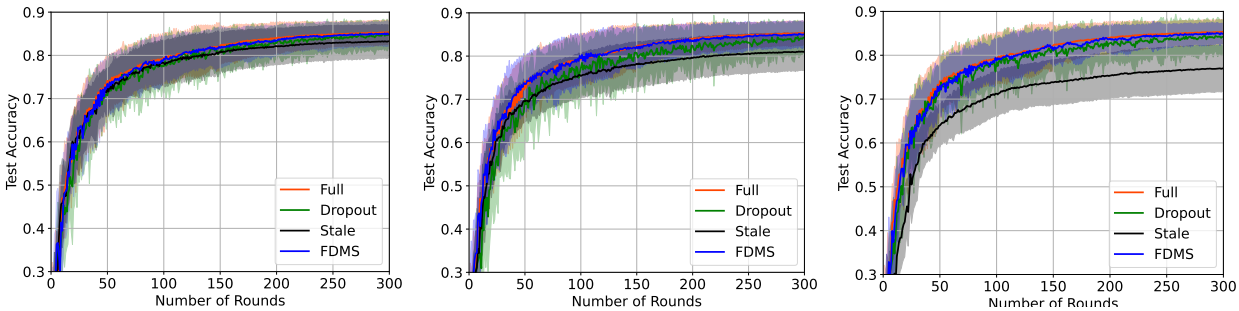
The results are presented in Fig. 6. We observed that as the level of non-i.i.d.-ness increased, our method **FL-FDMS** became less effective compared to **Full**, likely due to the increased difficulty in finding friends to perform the model substitution. Nonetheless, **FL-FDMS** still outperformed both the **Dropout** and **Stale** benchmarks, demonstrating that model substitution can enhance convergence compared to doing nothing or using an outdated model.

C.4. Experiments on the FMNIST datasets

We present additional performance comparison results in the FMNIST clustered setting, and the results in Fig. 7 are consistent with the conclusions we have drawn from prior other datasets.

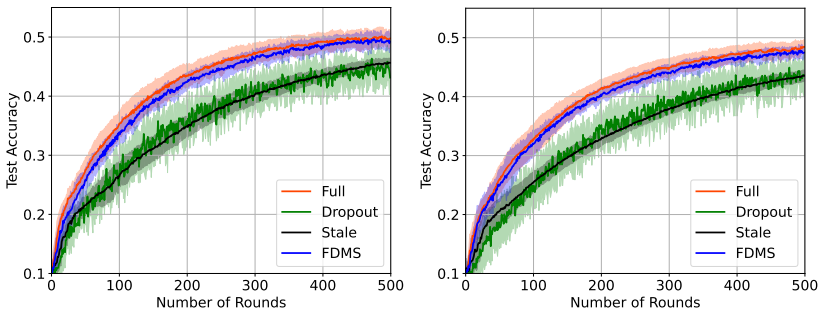
C.5. Experiments with FedProx

In the Cifar10 clustered setting, we have conducted additional performance comparison experiments using FedProx [21] with the FedProx parameter value $\mu = 0.2$. Our results, shown in Fig. 8, demonstrate that the proposed **FL-FDMS** algorithm remains effective.



(a) FMNIST ($\alpha = 0.3$) (b) FMNIST ($\alpha = 0.5$) (c) FMNIST ($\alpha = 0.7$)

Fig. 7: Performance comparison on the FMNIST clustered setting with various α



(a) CIFAR-10 (FedAvg) (b) CIFAR-10 (FedProx)

Fig. 8: Performance comparison on the CIFAR-10 clustered setting ($\alpha = 0.5$) with different FL algorithms