Appendix

A. Proof of Theorem 1

Proof. $\left\|F(\theta^{t})-F(\theta^{*})\right\|^{2}$

$= \left\|F(\theta^{t-1})-\eta\nabla F(\theta^{t-1})-F(\theta^{*})\right\|^{2}$

$= \left\|F(\theta^{t-1})-F(\theta^{*})\right\|^{2} - 2\eta\nabla F(\theta^{t-1})^{\mathrm{T}}(F(\theta^{t-1}) - F(\theta^{*})) + \eta^{2}\left\|\nabla F(\theta^{t-1})\right\|^{2}$

$\leq \left\|F(\theta^{t-1})-F(\theta^{*})\right\|^{2} - \eta\frac{\left\|\nabla\theta^{t-1}\right\|^{2}}{\beta} + \eta^{2}\left\|\nabla F(\theta^{t-1})\right\|^{2}$

$= \left\|F(\theta^{t-1})-F(\theta^{*})\right\|^{2} - \eta(\frac{1}{\beta}-\eta)\left\|\nabla F(\theta^{t-1})\right\|^{2}$

Finally, $\left\|F(\theta^{t})-F(\theta^{*})\right\|^{2} \leq \left\|F(\theta^{t-1})-F(\theta^{*})\right\|^{2}$, which ends the proof. □

B. Proof of Theorem 2

Proof. We prove it by induction. First, we list

$$\bar{\theta}^{1} = \theta^{0} - \eta\nabla\bar{g}^{1} \tag{1}$$

$$\theta_{l}^{1} = \theta^{0} - \eta\nabla g_{l}^{1} \tag{2}$$

where $\bar{g}^{1}$ and g_{l}^{1} denote the gradients of the FedAvg and *FBLG* model updates on the server at 1-th round, respectively. Using SGD for FedAvg in Eq. (1) and *FBLG* in Eq. (2) at 1-th round, we can easily draw the conclusion

$$\mathbb{E}\left\|\theta_{l}^{1}-\theta_{l}^{*}\right\|^{2} \leq \mathbb{E}\left\|\bar{\theta}^{1}-\theta_{l}^{*}\right\|^{2} \tag{3}$$

Then, assuming that Eq. (4) is established at t-th round, we can get

$$\mathbb{E}\left\|\theta_{l}^{t}-\theta_{l}^{*}\right\|^{2} \leq \mathbb{E}\left\|\bar{\theta}^{t}-\theta_{l}^{*}\right\|^{2} \tag{4}$$

Next, we use Eq. (3) and Eq. (4) to verify the $(t+1)$-th round, and then, we can get

$$\mathbb{E}\left\|\theta_{l}^{t}-\eta\nabla\bar{g}^{t}-\theta_{l}^{*}\right\|^{2} \leq \mathbb{E}\left\|\bar{\theta}^{t}-\eta\nabla g_{l}^{t}-\theta_{l}^{*}\right\|^{2} \tag{5}$$

which can be further expressed as

$$\mathbb{E}\left\|\theta_{l}^{t+1}-\theta_{l}^{*}\right\|^{2} \leq \mathbb{E}\left\|\bar{\theta}^{t+1}-\theta_{l}^{*}\right\|^{2} \tag{6}$$

Thus, we end the proof. □

C. Experimental Details

Data Visualization We use $\alpha = \{0.8,\ 0.05,\ 0.01\}$ in our experiments and visualize the partition of each dataset for the slightly skewed case (*i.e.* $\alpha = 0.8$) and the extremely skewed case (*i.e.* $\alpha = 0.01$), as shown in Fig. 1. We observe that almost only a small fraction of the clients have two different labels and the rest of the clients have only one label on the four datasets at extremely skewed data partitioning (*i.e.* $\alpha = 0.01$), while almost all the clients have multiple labels on the four datasets at slightly skewed data partitioning (*i.e.* $\alpha = 0.8$).

Results of Test Loss We plot the testing loss curves of our proposed *FBLG* and 9 baseline methods respectively on the FMNIST and CIFAR10 datasets under the second degree of skew (*i.e.* $\alpha = 0.05$), as shown in Fig. 2. We can observe that *FBLG* converges stably and gradually tends to the optimal.

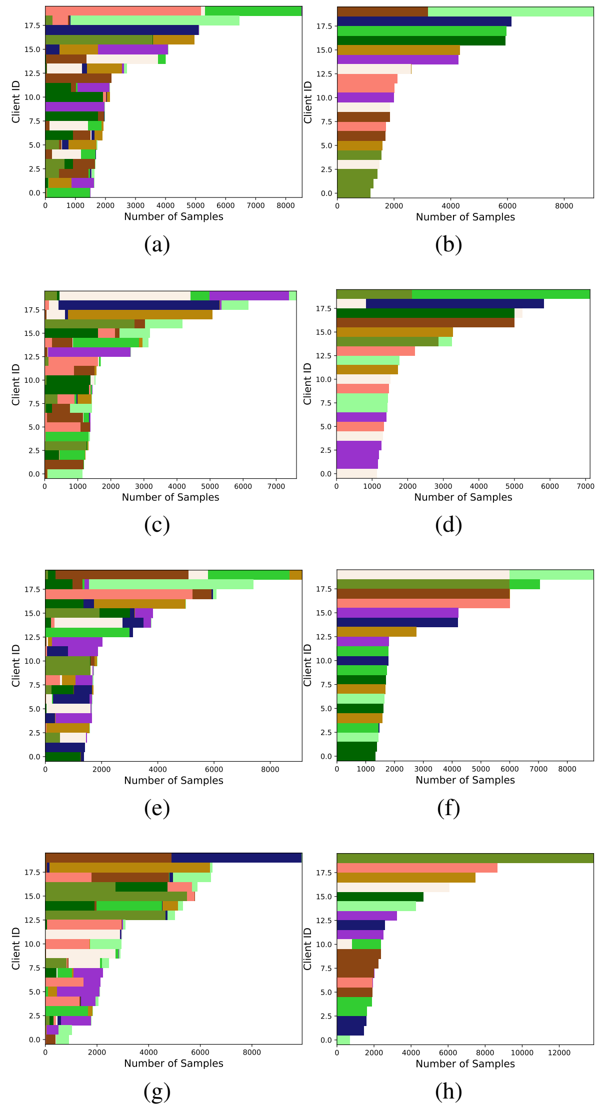

Figure 1: The visualization of skewed data respectively for the four datasets MNIST, CIFAR10, FMNIST, and SVHN from top to bottom, where $\alpha = 0.8$ for the left and $\alpha = 0.01$ for the right.

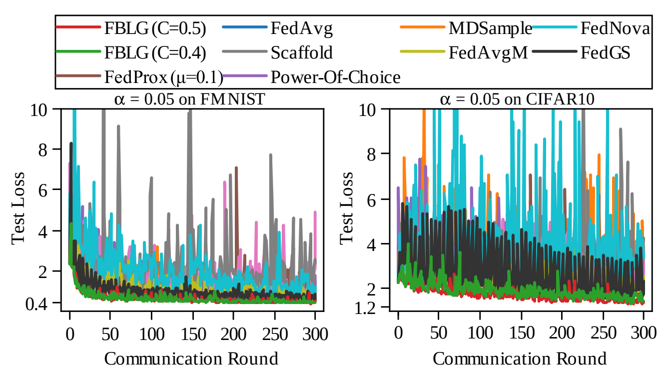

Figure 2: Testing loss curves respectively on FMNIST and CIFAR10 when $\alpha = 0.05$.