A Details of dataset and experimental setups

We implement the proposed scaled delta and advanced free-rider attacks in PyTorch. Our simulated FL environment runs on our personal computer. The main function works as the central server. The participants include 50 honest clients and a certain number of free-riders including 1, 50 and 150. All clients conduct local training for E epochs over their local datasets and then return local updates by using API. The central server then conducts FedAvg algorithm and distribute the aggregated model in next round. As for models, we use a MLP model for MNIST dataset and CNN model for Fashion-MNIST dataset.

MNIST . For the case where n clients participate in the training process, we assign 60000 samples to training and 10000 samples to testing and create two settings: an iid dataset (MNIST iid) where each client randomly selects $60000/n$ samples from the whole data set as the local data set without replacement, and a non-iid dataset (MNIST non-iid), where for each digit we create two shards with $60000/2n$ training samples, and allocate 2 shards for each client.

Fashion-MNIST . We study a CNN model for classification problem on the dataset of Fashion-MNIST. Here we adpot the same data division scheme as that of MNIST. Detailed information of the CNN architecture are summarized in TableA

We test federated learning with 5 local epochs using SGD optimization with learning rate $\gamma = 0.001$ for MNIST (iid and non-iid), and $\gamma = 0.002$ for Fashion-MNIST , and batch size of 128. All clients participate in each round of training process.

Algorithm 2: The datasets used and their associated learning models and hyper-parameters.

Parameter	Shape Layer	hyper-parameter
layer1.conv1.weight	$1 \times 20 \times 5 \times 5$	stride:1, padding:0
layer1.conv1.bias	20	N/A
pooling.max	N/A	kernel size:2;stride:2
layer2.conv2.weight	$20 \times 50 \times 5 \times 5$	stride:1;padding:0
layer2.conv2.bias	50	N/A
pooling.max	N/A	kernel size:2;stride:2
layer3.fc1.weight	800×500	N/A
layer3.fc1.bias	500	N/A
layer4.fc2.weight	500×10	N/A
layer4.fc2.bias	10	N/A

B Additional experimental results

B.1 Experiments for non-iid setting of MNIST with MLP

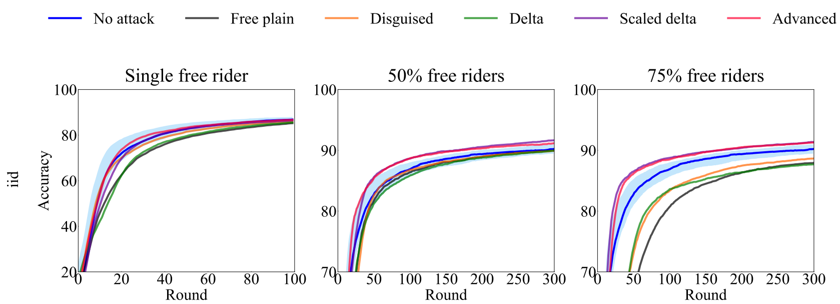


Figure 4: Accuracy performance in non-iid setting of MNIST dataset

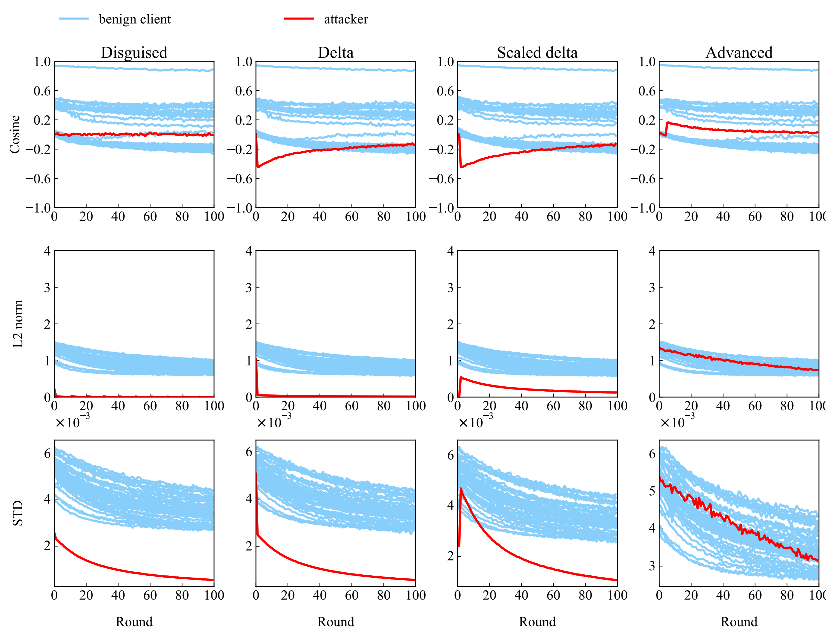


Figure 5: Features in non-iid setting of MNIST dataset

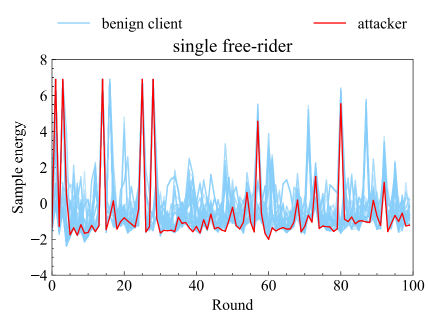


Figure 6: Sample energy in non-iid setting of MNIST dataset

B.2 Experiments for Fashion-MNIST with CNN

B.2.1 Convergence and Performances.

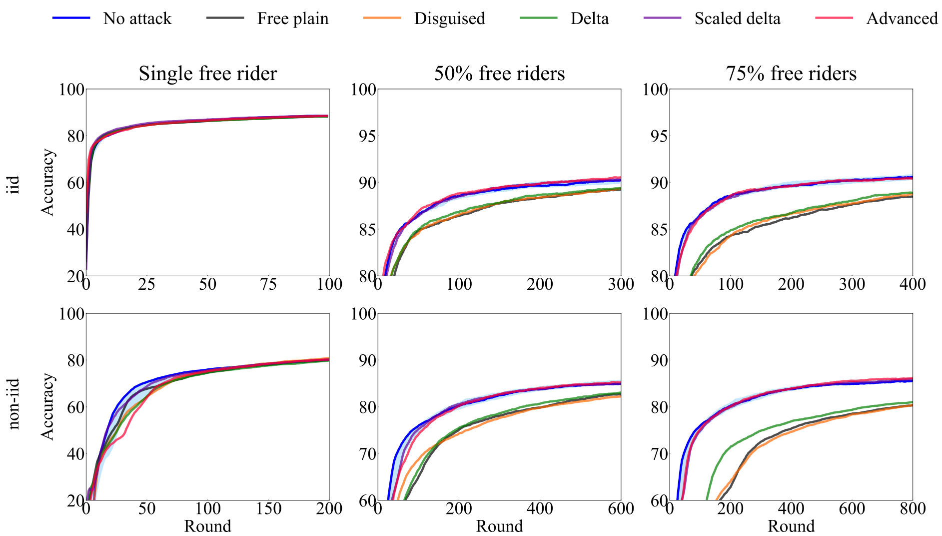


Figure 7: Accuracy performance of Fashion-MNIST dataset

B.2.2 Stealth Property.

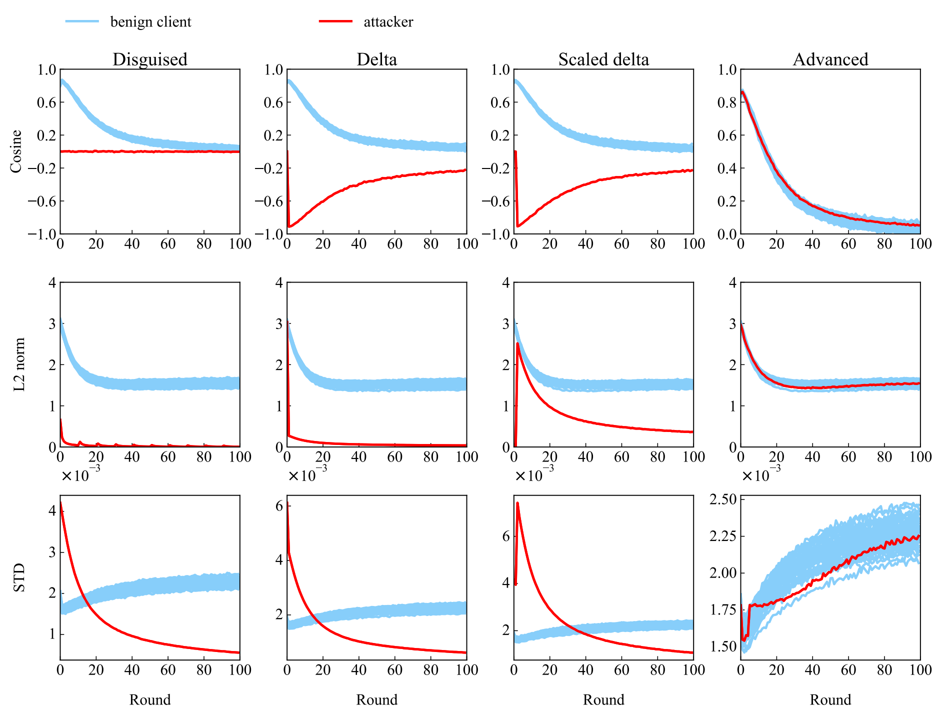


Figure 8: Features in iid setting of Fashion-MNIST dataset

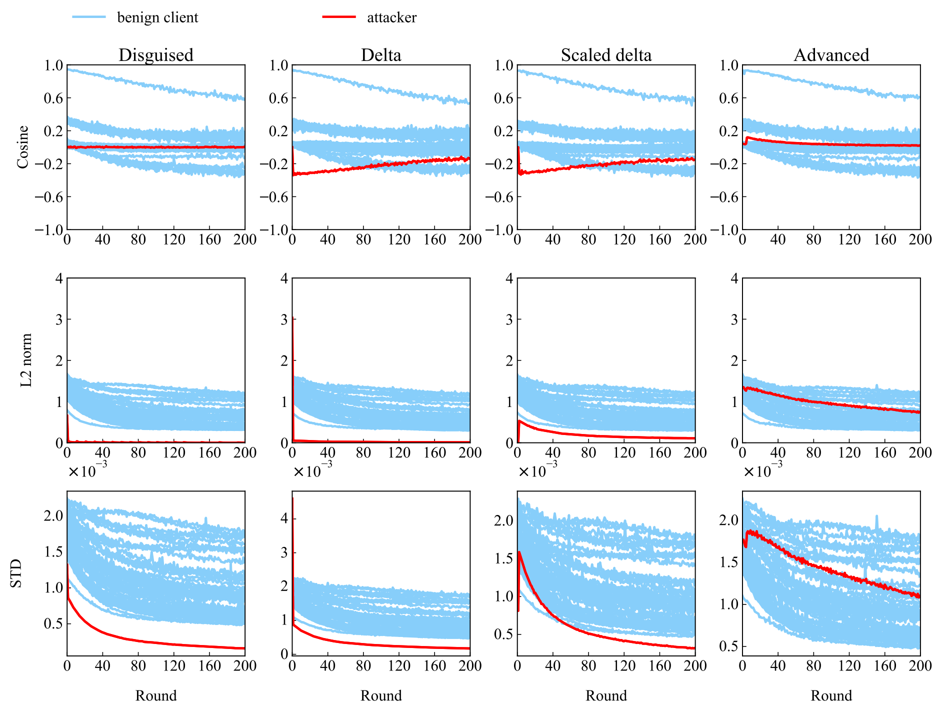
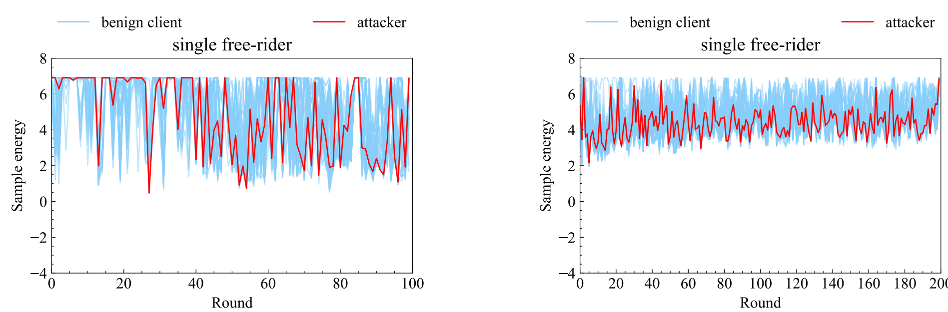


Figure 9: Features in non-iid setting of Fashion-MNIST dataset

B.2.3 Anomaly Detection.



(a) iid

(b) non-iid

Figure 10: Sample energy for Fashion-MNIST dataset

C Complete Proofs

C.1 Proof of Theorem 1

Proof. The continuous-time stochastic differential equation is:

$$d\theta_i = -g_i(\theta)dt + \frac{\gamma}{\sqrt{s}}BdW_t, \tag{12}$$

We can replace the parameter $g_i(\theta)$ in equation 12 according to assump 1. Then we can have

$$d\theta_i \approx \lambda_i(\theta_i - \theta_i(t))dt + \frac{\gamma}{\sqrt{s}}BdW_t. \tag{13}$$

The differential equation is solved:

$$\theta_i(t) = e^{-\lambda_i t}[C_0 + \underbrace{\int_0^t \lambda_i\theta_i e^{\lambda_i t}dt}_{A} + \underbrace{\frac{\gamma}{\sqrt{s}}\int_0^t B(t)e^{\lambda_i t}dW_t]}_{B} \tag{14}$$

In the initial round, the client receives the global model $\theta(0)$ as its local model. Thus, we have $\theta_i(0) = \theta(0)$ and $C_0 = \theta(0)$.

We study (A) :

$$\int_0^t \lambda_i\theta_i e^{\lambda_i t}dt = (e^{\lambda_i t} - 1)\theta_i. \tag{15}$$

Moreover, we could obtain a specific equation which describes the evolution of local training model weights. The solution of this equation is

$$\theta_i(t) \approx \theta(0)e^{-\lambda_i t} + (1 - e^{-\lambda_i t})\theta_i + \frac{\gamma}{\sqrt{s}}e^{-\lambda_i t}\int_0^t e^{\lambda_i t}BdW_t. \tag{16}$$

Asymptotic convergence of B The asymptotic properties of the stochastic integral B follows from the general properties of Itos integrals. For any constant L, we have

$$\mathbb{E}[Le^{-\lambda_i t}\int_0^t e^{\lambda_i t}dW_t] \xrightarrow{t\to+\infty} 0 \tag{17}$$

$$Var[Le^{-\lambda_i t}\int_0^t e^{\lambda_i t}dW_t] = L^2e^{-2\lambda_i t}\int_0^t e^{2\lambda_i t}dW_t = \frac{L^2}{2\lambda_i}(1 - e^{-2\lambda_i t}) \xrightarrow{t\to+\infty} \frac{L^2}{2\lambda_i}. \tag{18}$$

According to Eq. 14, we finally conclude that

$$\begin{aligned} \mathbb{E}[\theta_i(t)] &\xrightarrow{t\to+\infty} \theta_i, \\ Var[\theta_i(t)] &\xrightarrow{t\to+\infty} \frac{\gamma^2B^2}{2\lambda_i s}. \end{aligned} \tag{19}$$

□

C.2 Proof of Theorem 2

Proof. The aggregation at the server level follows:

$$\theta(t+1) = \theta(t) + \sum_{i=1}^{n}\frac{N_i}{N}U_i(\theta). \tag{20}$$

As illustrated in Section3.2, the local model update for client i follows:

$$U_i(\theta) \approx d\theta_i = -g_i(\theta)dt + \frac{\gamma}{\sqrt{s}}B_i dW_t. \tag{21}$$

Similar to the discretization equation and continuous-time stochastic differential equation illustrated in Eq. 12, we could further get

$$\begin{aligned}
\theta(t+1)-\theta(t) &= \sum_{i=1}^{n}\frac{N_i}{N}U_i(\theta), \\
d\theta(t) &\approx -\sum_{i=1}^{n}\frac{N_i}{N}g_i(\theta)dt+\sum_{i=1}^{n}\frac{N_i}{N}\frac{\gamma}{\sqrt{s}}B_i dW_t \\
&= \sum_{i=1}^{n}\frac{N_i\lambda_i}{N}(\theta_i-\theta_i(t))dt+\sum_{i=1}^{n}\frac{N_i}{N}\frac{\gamma}{\sqrt{s}}B_i dW_t.
\end{aligned} \tag{22}$$

Client C_i receives the global model $\theta(t)$ as initial model parameters from the central server at round t, that is, $\theta_i(t)=\theta(t)$. We define $\overline{\lambda}=\sum_{i=1}^{n}\frac{N_i\lambda_i}{N}$ and $\overline{\theta}=\frac{\sum_{i=1}^{n}N_i\lambda_i\theta_i}{\sum_{i=1}^{n}N_i\lambda_i}$. And we have

$$d\theta(t)\approx\overline{\lambda}(\overline{\theta}-\theta(t))+\sum_{i=1}^{n}\frac{N_i}{N}\frac{\gamma}{\sqrt{s}}B_i dW_t. \tag{23}$$

We note that this equation has the same form with the one in proof C.1. The solution is

$$\theta(t)\approx\theta(0)e^{-\overline{\lambda}t}+(1-e^{-\overline{\lambda}t})\overline{\theta}+\sum_{i=1}^{n}\frac{N_i}{N}\frac{\gamma}{\sqrt{s}}e^{-\overline{\lambda}t}\int_0^t B_i e^{\overline{\lambda}t}dW_t. \tag{24}$$

$$\begin{aligned}
\mathbb{E}[\theta(t)] &\xrightarrow{t\to+\infty}\overline{\theta}, \\
Var[\theta(t)] &\xrightarrow{t\to+\infty}\sum_{i=1}^{n}\frac{\gamma^2N_i^2B_i^2}{2\overline{\lambda}N^2s}.
\end{aligned} \tag{25}$$

□

C.3 Proof of Theorem 3

Proof. We denote by J_1 the set of honest clients and by J_2 the set of m free-riders. The total number of training samples declared by free_riders is M. At round t, the aggregated model with free-riders could be calculated as follows:

$$\begin{aligned}
\theta(t+1) &= \theta(t)+\sum_{J_1\cup J_2}\frac{N_i}{N}U_i(\theta) \\
&= \theta(t)+\sum_{i\in J_1}\frac{N_i}{N}U_i(\theta)+\sum_{i\in J_2}\frac{N_i}{N}U_f(\theta) \\
&= \theta(t)+\sum_{i\in J_1}\frac{N_i}{N}U_i(\theta)+\frac{M}{N}(\frac{\|\theta(t)-\theta(t-1)\|}{\|\theta(t-1)-\theta(t-2)\|}(\theta(t)-\theta(t-1)))
\end{aligned} \tag{26}$$

According to the discussion in Section 4.1, we could obtain

$$\theta(t+1)-\theta(t)\approx\frac{\|\theta(t)-\theta(t-1)\|}{\|\theta(t-1)-\theta(t-2)\|}(\theta(t)-\theta(t-1)) \tag{27}$$

Rearrange this equation, we could have

$$\begin{aligned}
\theta(t+1)-\theta(t) &= \sum_{i\in J_1}\frac{N_i}{N-M}U_i(\theta), \\
d\theta(t) &\approx -\sum_{i\in J_1}\frac{N_i}{N-M}g_i(\theta)dt+\sum_{i\in J_1}\frac{N_i}{N-M}\frac{\gamma}{\sqrt{s}}B_i dW_t. \\
&= \sum_{i\in J_1}\frac{N_i}{N-M}\lambda_i(\theta_i-\theta_i(t))dt+\sum_{i\in J_1}\frac{N_i}{N-M}\frac{\gamma}{\sqrt{s}}B_i dW_t.
\end{aligned} \tag{28}$$

We note that this equation has the same form with the one in proof C.2 The solution is

$$\theta(t) \approx \theta(0)e^{-\overline{\lambda}t} + (1 - e^{-\overline{\lambda}t})\overline{\theta} + \sum_{i \in J_1} \frac{N_i}{N-M} \frac{\gamma}{\sqrt{s}} e^{-\overline{\lambda}t} \int_0^t B_i e^{\overline{\lambda}t} dW_t. \tag{29}$$

$$\begin{aligned} \mathbb{E}[\theta(t)] &\xrightarrow{t \to +\infty} \overline{\theta}, \\ Var[\theta(t)] &\xrightarrow{t \to +\infty} \sum_{i \in J_1} \frac{\gamma^2 N_i^2 B_i^2}{2\overline{\lambda}(N-M)^2 s}, \end{aligned} \tag{30}$$

where $\overline{\lambda} = \sum_{i \in J_1} \frac{N_i}{N-M} \lambda_i$ and $\overline{\theta} = \frac{\sum_{i \in J_1} N_i \lambda_i \theta_i}{\sum_{i \in J_1} N_i \lambda_i}$ □

C.4 Proof of Lemma 1

Proof. According to the discussion in section 3.2 and our assumption 1, at round t, the local model $\theta_i(t)$ is initiated with the global model $\theta(t)$, thus we could get

$$\begin{aligned} \mathbb{E}(U_i(\theta)) &\approx \mathbb{E}(d\theta_i) \\ &= \mathbb{E}(-g_i(\theta)dt + \frac{\gamma}{\sqrt{s}} B_i dW_t) \\ &= \gamma\lambda(\theta_i - \theta(t)). \end{aligned} \tag{31}$$

As shown in Eq. 24, we could infer that

$$\begin{aligned} \mathbb{E}(\theta_i - \theta(t)) &\approx \mathbb{E}(\theta_i - \theta(0)e^{-\overline{\lambda}t} - (1 - e^{-\overline{\lambda}t})\overline{\theta}) \\ &= \mathbb{E}(\theta_i - \overline{\theta} + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}) \\ &= \mathbb{E}(\epsilon_i + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}) \end{aligned} \tag{32}$$

where we define $\epsilon_i = \theta_i - \overline{\theta}$. Then we could compute the expection of cosine value β for random two local updates $U_i(\theta)$ and $U_j(\theta)$ as follows:

$$\begin{aligned} \mathbb{E}(\cos\beta) &= \frac{\mathbb{E}(U_i(t)) \cdot \mathbb{E}(U_j(t))}{|\mathbb{E}(U_i(t))| \cdot |\mathbb{E}(U_j(t))|} \\ &\approx \frac{(\epsilon_i + (\overline{\theta} - \theta(0))e^{\overline{\lambda}t}) \cdot (\epsilon_j + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t})}{\left|(\epsilon_i + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}\right| \cdot \left|(\epsilon_j + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}\right|}. \\ &= \frac{\epsilon_i \cdot \epsilon_j + (\epsilon_i + \epsilon_j) \cdot (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t} + ((\overline{\theta} - \theta(0))e^{-\overline{\lambda}t})^2}{\left|(\epsilon_i + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}\right| \left|(\epsilon_j + (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t}\right|}. \end{aligned} \tag{33}$$

As illustrated in Section 4.2, ϵ_i, ϵ_j are two random values which demonstrate the difference between local client models and global model, which is resulted from the difference between local datasets and whole dataset. According to the discussion in [26], they provided a precise characterization of the folklore that all high-dimensional random vectors are almost always nearly orthogonal to each other. Therefore, we could have

$$\epsilon_i \cdot \epsilon_j = 0. \tag{34}$$

$\theta(0)$ is the random initialization parameter of the model. Similarly, we can get

$$(\epsilon_i + \epsilon_j) \cdot (\overline{\theta} - \theta(0))e^{-\overline{\lambda}t} = 0. \tag{35}$$

We assume that $|\mathbb{E}(\overline{\theta}-\theta(0))|= C|\mathbb{E}(\epsilon_i)|$, where C is an introduced parameter. Finally we could conclude that:

$$\begin{aligned}
\mathbb{E}(\cos\beta) &= \frac{\epsilon_i\cdot\epsilon_j + (\epsilon_i+\epsilon_j)\cdot(\overline{\theta}-\theta(0))e^{-\overline{\lambda}t} + ((\overline{\theta}-\theta(0))e^{-\overline{\lambda}t})^2}{\left|(\epsilon_i + (\overline{\theta}-\theta(0))e^{-\overline{\lambda}t}\right|\left|(\epsilon_j + (\overline{\theta}-\theta(0))e^{-\overline{\lambda}t}\right|} \\
&= \frac{(\overline{\theta}-\theta(0))^2 e^{-2\overline{\lambda}t}}{\epsilon^2 + (\overline{\theta}-\theta(0))^2 e^{-2\overline{\lambda}t}} \\
&= \frac{C^2}{C^2 + e^{2\overline{\lambda}t}}.
\end{aligned} \tag{36}$$

□

C.5 Proof of Lemma 2

Proof. According to previous Eq. 20, Eq. 31 and Eq. 32, we now could calculate the *l2*-norm of expecation for global update $\theta(t)-\theta(t-1)$ as follows:

$$\begin{aligned}
|\mathbb{E}(\theta(t)-\theta(t-1))| &= |\mathbb{E}(-\sum_{i=1}^{n}\frac{N_i}{N}U_i(\theta))| \\
&\approx |\mathbb{E}(\sum_{i=1}^{n}\frac{N_i}{N}\gamma\lambda_i(\theta_i-\theta(t-1)))|
\end{aligned} \tag{37}$$

For iid setting, we assume that λ_i is the same for all clients, which is equal to λ. The expected number of local training samples declared by client C_i is equal to N/n,that is,$E(N_i)$. The expectation l2-norm value of the local model update returned by client C_i at round t is $|U|$, i.e. $|\mathbb{E}(U_i(\theta))|= |U|$,So

$$\begin{aligned}
|\mathbb{E}(\theta(t)-\theta(t-1))| &= \frac{\gamma\lambda}{n}|\mathbb{E}(\sum_{i=1}^{n}(\theta_i-\theta(t-1))| \\
&= \frac{\gamma\lambda}{n}\sqrt{\mathbb{E}((\sum_{i=1}^{n}(\theta_i-\theta(t-1)))^2)} \\
&= \frac{\gamma\lambda}{n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(\theta_i-\theta(t-1))\mathbb{E}(\theta_j-\theta(t-1))} \\
&= \frac{1}{n}\sqrt{n+(n^2-n)\mathbb{E}(\cos\beta)}|U|.
\end{aligned} \tag{38}$$

Thus we could compute

$$\frac{|\mathbb{E}(U_i(\theta))|}{|\mathbb{E}(\theta(t)-\theta(t-1))|} = \sqrt{\frac{n^2}{n+(n^2-n)\mathbb{E}(\cos\beta)}}. \tag{39}$$

According to the previous analysis 4.1, the weighted average of all local updates $\overline{U}(\theta) = \theta(t) - \theta(t-1)$ could be viewed as a global update over the whole training dataset with a one-step training, by assumption 1. We could get $\overline{U}(\theta) = \gamma\lambda(\overline{\theta}-\theta(t-1))$. We could compute the difference between $\overline{U}(\theta)$ and a real local update $U_i(\theta)$ as follows.

$$\begin{aligned}
\mathbb{E}(U_i(\theta)-\overline{U}(\theta)) &\approx \mathbb{E}(\gamma\lambda(\theta_i-\theta(t-1)) - \gamma\lambda(\overline{\theta}-\theta(t-1))) \\
&= \mathbb{E}(\gamma\lambda(\theta_i-\overline{\theta})) \\
&= \mathbb{E}(\gamma\lambda\epsilon_i).
\end{aligned} \tag{40}$$

Then, we could obtain:

$$\mathbb{E}(U_i(\theta)) = \mathbb{E}(\overline{U}(\theta)) + \mathbb{E}(\gamma\lambda\epsilon_i). \tag{41}$$

According to the Eq. 24, we have $\mathbb{E}(\overline{U}(\theta)) = (\bar{\theta}-\theta(0))(e^{-(\lambda-1)t} - e^{-\lambda t})$. In Eq.35, the vector $\bar{\theta}-\theta(0)$ is orthogonal to vector ϵ_i, that is, $\overline{U}(\theta)$ is orthogonal to vector ϵ_i. □

C.6 Proof of Theorem 4

Proof. Similar to the proof of Theorem 3, we could obtain

$$\begin{aligned}\theta(t+1) &= \theta(t) + \sum_{i=1}^{n} \frac{N_i}{N} U_i(\theta) \\ &= \theta(t) + \sum_{i\in J_1} \frac{N_i}{N} U_i(\theta) + \frac{M}{N} \sum_{i\in J_2} \hat{U}_f(\theta) \\ &= \theta(t) + \sum_{i\in J_1} \frac{N_i}{N} U_i(\theta) + \frac{M}{N} \left(\frac{\|\theta(t)-\theta(t-1)\|}{\|\theta(t-1)-\theta(t-2)\|} (\theta(t)-\theta(t-1)) + \varphi(t)\right) dW_{t_2} \\ &= \theta(t) + \sum_{i\in J_1} \frac{N_i}{N} U_i(\theta) + \frac{M}{N} \frac{\|\theta(t)-\theta(t-1)\|}{\|\theta(t-1)-\theta(t-2)\|} (\theta(t)-\theta(t-1)) + \frac{M}{N} \varphi(t) dW_{t_2}.\end{aligned} \tag{42}$$

Rearrange this equation, we could have

$$\begin{aligned}\theta(t+1)-\theta(t) &= \sum_{i\in J_1} \frac{N_i}{N-M} U_i(\theta) + M\varphi(t) dW_{t_2} \\ d\theta(t) &\approx \sum_{i\in J_1} \frac{N_i}{N-M} U_i(\theta) + M\varphi(t) dW_{t_2} \\ &= -\sum_{i\in J_1} \frac{N_i}{N-M} g_i(\theta) dt + \sum_{i\in J_1} \frac{N_i}{N-M} \frac{\sqrt{\gamma}}{s} B_i dW_{t_1} + \frac{M}{N-M} \varphi(t) dW_{t_2}.\end{aligned} \tag{43}$$

According to our assumption 1, we know that

$$d\theta(t) \approx \sum_{i\in J_1} \frac{N_i}{N-M} \lambda_i(\theta_i - \theta_i(t)) dt + \sum_{i\in J_1} \frac{N_i}{N-M} \frac{\gamma}{\sqrt{s}} B_i dW_{t_1} + \frac{M}{N-M} \varphi(t) dW_{t_2}. \tag{44}$$

The solution for this equation is

$$\theta(t) \approx \theta(0) e^{-\overline{\lambda} t} + (1 - e^{-\overline{\lambda} t})\overline{\theta} + \sum_{i\in J_1} \frac{N_i}{N-M} \frac{\gamma}{\sqrt{s}} e^{-\overline{\lambda} t} \int_0^t B_i e^{\overline{\lambda} t} dW_{t_1} + \underbrace{\frac{M}{N-M} e^{-\overline{\lambda} t} \int_0^t \varphi(t) e^{\lambda t} dW_{t_2}}_{A}. \tag{45}$$

We note that this equation differs from the one in Eq.28 for the last term only.

We study (A):

Because $|\mathbb{E}(\varphi(t))| = \sqrt{\frac{n^2}{n+(n^2-n)\mathbb{E}(\cos\beta)} - 1}\,|\mathbb{E}(U_f(\theta))|$, we know that $(U_f(\theta))|$ decays at a rate of $O(e^{-\overline{\lambda} t})$ and $\sqrt{\frac{n^2}{n+(n^2-n)\mathbb{E}(\cos\beta)} - 1}$ is bounded. There exists a constant p which satisfies $|\mathbb{E}(\varphi(t))| = \sqrt{\frac{n^2}{n+(n^2-n)\mathbb{E}(\cos\beta)} - 1}\,|\mathbb{E}(U_f(\theta))| \leq p e^{-\overline{\lambda} t}$, then we could obtain

$$Var\left[\frac{M}{N-M} e^{-\overline{\lambda} t} \int_0^t \varphi(t) e^{\overline{\lambda} t} dW_{t_2}\right] \leq \frac{M^2 p^2}{(N-M)^2} e^{-2\overline{\lambda} t} \int_0^t dW_{t_2} \tag{46}$$

$$\frac{M^2 p^2}{(N-M)^2} e^{-2\overline{\lambda} t} \int_0^t dW_{t_2} \xrightarrow{t\to+\infty} 0 \tag{47}$$

Similarly, we could obtain the expectation and variation value for the global model as follows:

$$\begin{aligned} \mathbb{E}[\theta(t)] &\xrightarrow{t\to+\infty} \tilde{\theta}, \\ Var[\theta(t)] &\xrightarrow{t\to+\infty} \sum_{i\in J_1} \frac{\gamma^2 N_i^2 B_i^2}{2\overline{\lambda}(N-M)^2 s}, \end{aligned} \tag{48}$$

where $\overline{\lambda} = \sum_{i\in J_1} \frac{N_i}{N-M}\lambda_i$ and $\overline{\theta} = \frac{\sum_{i\in J_1} N_i\lambda_i\theta_i}{\sum_{i\in J_1} N_i\lambda_i}$. We could infer that the expectation of the global model $\theta(t)$ is the average of the expecation of all local model updates from honest clients, and the variation value is $\sum_{i\in J_1} \frac{\gamma^2 N_i^2 B_i^2}{2\overline{\lambda}(N-M)^2 s}$. Therefore, our advanced free-rider attack also guarantees the convergence property. □