FedCS: Coreset Selection for Federated Learning

Supplementary Material

A. Preliminaries for Proof of Theorem 4.1

Notation: Necessary notations are introduced as follows.

- $\mathbf{w}_i^{(t)}$: local weight vector on client c_i at epoch t.
- D_i^*: coreset on client c_i after pruning.
- $D^* = \bigcup_{i=1}^N D_i$
- $\mathbf{g}_i(\mathbf{w}_i^{(t)}, \xi_i^{(t)}) = \nabla F_i(\mathbf{w}_i, D_i^*)$
- $\mathbf{w}^{(t)} = \frac{1}{m}\sum_{i\in\mathcal{S}} \mathbf{w}_i^{(t)}$

We present the preliminary lemmas used for proof of Theorem 4.1. We will denote the expectation over the sampling random source $\mathcal{S}(t)$ as $\mathbb{E}_{\mathcal{S}(t)}$ and the expectation over all the random sources as $\mathbb{E}$.

Lemma A.1. *Suppose F_i is L-smooth with global minimum at $\mathbf{w}_i^*$, then for any $\mathbf{w}_i$ in the domain of F_i, we have that*

$$\|\nabla F_i(\mathbf{w}_i, D_i^*)\|^2 \leq 2L(F_i(\mathbf{w}_i, D_i^*) - F_i(\mathbf{w}_i^*, D_i^*)). \quad (18)$$

Proof.

$$F_i(\mathbf{w}_i, D_i^*) - F_i(\mathbf{w}_i^*, D_i^*) - < \nabla F_i(\mathbf{w}_i^*, D_i^*), \mathbf{w}_i - \mathbf{w}_i^* > \geq \frac{1}{2L}\|\nabla F_i(\mathbf{w}_i, D_i^*) - \nabla F_i(\mathbf{w}_i^*, D_i^*)\|^2. \quad (19)$$

$$F_i(\mathbf{w}_i, D_i^*) - F_i(\mathbf{w}_i^*, D_i^*) \geq \frac{1}{2L}\|\nabla F_i(\mathbf{w}_i, D_i^*)\|^2. \quad (20)$$

Lemma A.2 (Expected average discrepancy between $\mathbf{w}^{(t)}$ and $\mathbf{w}_i^{(t)}$ for $i \in \mathcal{S}(t)$)**.**

$$\frac{1}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)} \|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2] \leq 16\eta_t^2\tau^2G^2. \quad (21)$$

Proof.

$$\begin{aligned}
&\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \\
&= \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \|\frac{1}{m}\sum_{i'\in\mathcal{S}(t)} (\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)})\|^2 \\
&\leq \frac{1}{m^2}\sum_{i\in\mathcal{S}(t)}\sum_{i'\in\mathcal{S}(t)} \|(\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)})\|^2 \\
&= \frac{1}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \|(\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)})\|^2. \quad (22)
\end{aligned}$$

Moreover, for any t, there is a t_0 such that $\mathbf{w}_{i'}^{(t_0)} = \mathbf{w}_i^{(t_0)}$ and $0 \leq t - t_0 < \tau$, because the selected clients are updated with the global model at every τ . Hence, even for an arbitrary t, we have the difference $\|\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)}\|^2$ is upper bounded by τ updates. With non-increasing η_t over t and $\eta_{t_0} \leq 2\eta_t$, equation (22) can be further bounded as,

$$\begin{aligned}
&\frac{1}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \|\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \\
&\leq \frac{1}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \|\sum_{j=t_0}^{t_0+\tau-1} \eta_j(\mathbf{g}_{i'}(\mathbf{w}_{i'}^{(j)}, \xi_{i'}^{(j)}) - \\
&\quad \mathbf{g}_i(\mathbf{w}_i^{(j)}, \xi_i^{(j)}))\|^2 \\
&\leq \frac{\eta_{t_0}^2\tau}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \sum_{j=t_0}^{t_0+\tau-1} \|\mathbf{g}_{i'}(\mathbf{w}_{i'}^{(j)}, \xi_{i'}^{(j)}) - \mathbf{g}_i(\mathbf{w}_i^{(j)}, \xi_i^{(j)})\|^2 \\
&\leq \frac{\eta_{t_0}^2\tau}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \sum_{j=t_0}^{t_0+\tau-1} \Big[2\|\mathbf{g}_{i'}(\mathbf{w}_{i'}^{(j)}, \xi_{i'}^{(j)})\|^2 + \\
&\quad 2\|\mathbf{g}_i(\mathbf{w}_i^{(j)}, \xi_i^{(j)})\|^2\Big]. \quad (23)
\end{aligned}$$

By taking expectation over (23),

$$\begin{aligned}
&\mathbb{E}[\frac{1}{m^2}\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \|\mathbf{w}_{i'}^{(t)} - \mathbf{w}_i^{(t)}\|^2] \\
&\leq \frac{2\eta_{t_0}^2\tau}{m^2}\mathbb{E}[\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \sum_{j=t_0}^{t_0+\tau-1} (\|\mathbf{g}_{i'}(\mathbf{w}_{i'}^{(j)}, \xi_{i'}^{(j)})\|^2 \\
&\quad + \|\mathbf{g}_i(\mathbf{w}_i^{(j)}, \xi_i^{(j)})\|^2)] \\
&\leq \frac{2\eta_{t_0}^2\tau}{m^2}\mathbb{E}_{\mathcal{S}(t)}[\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} \sum_{j=t_0}^{t_0+\tau-1} 2G^2] \\
&= \frac{2\eta_{t_0}^2\tau}{m^2}\mathbb{E}_{\mathcal{S}(t)}[\sum_{\substack{i\neq i'\\ i,i'\in\mathcal{S}(t)}} 2\tau G^2] \\
&\leq \frac{16\eta_t^2(m-1)\tau^2G^2}{m} \leq 16\eta_t^2\tau^2G^2. \quad (24)
\end{aligned}$$

Lemma A.3 (Upper bound for expectation over $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|$)**.** *By using $\mathbb{E}[\cdot]$, we have the upper*

bound of the total expectation over all random sources as:

$$\mathbb{E}[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2] \leq \frac{1}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)} \|\mathbf{w}_i^{(t)} - \mathbf{w}^*\|^2]. \quad (25)$$

Proof.

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2] &= \mathbb{E}[\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \mathbf{w}_i^{(t)} - \mathbf{w}^*\|^2] \\ &= \mathbb{E}[\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} (\mathbf{w}_i^{(t)} - \mathbf{w}^*)\|^2] \\ &\leq \frac{1}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)} \|\mathbf{w}_i^{(t)} - \mathbf{w}^*\|^2]. \quad (26)\end{aligned}$$

B. Proof of Theorem 4.1

With $\overline{\mathbf{g}}^{(t)} = \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \mathbf{g}_i(\mathbf{w}_i^{(t)}, \xi_i^t)$, we have that:

$$\begin{aligned}&\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 = \|\mathbf{w}^{(t)} - \eta_t\overline{\mathbf{g}}^{(t)} - \mathbf{w}^*\|^2 \\ &= \|\mathbf{w}^{(t)} - \eta_t\overline{\mathbf{g}}^{(t)} - \mathbf{w}^* - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) \\ &\quad + \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2 \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2 \\ &\quad + \eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)}\|^2 \\ &\quad + 2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^* - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*), \\ &\quad \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)} > \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \\ &\quad - 2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^*, \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &\quad + 2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^* - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*), \\ &\quad \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)} > \\ &\quad + \eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2 \\ &\quad + \eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)}\|^2. \quad (27)\end{aligned}$$

Then we defined:

$$A_1 = -2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^*, \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) >, \quad (28)$$

$$\begin{aligned}A_2 =&2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^* - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*), \\ &\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)} >, \quad (29)\end{aligned}$$

$$A_3 = \eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2, \quad (30)$$

$$A_4 = \eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) - \overline{\mathbf{g}}^{(t)}\|^2. \quad (31)$$

First, to obtain the upper bound of A_1(28):

$$\begin{aligned}&- 2\eta_t < \mathbf{w}^{(t)} - \mathbf{w}^*, \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &= -\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &= -\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &\quad - \frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}_i^{(t)} - \mathbf{w}^*, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > . \quad (32)\end{aligned}$$

By using the Cauchy inequality and AM-GM inequality, we can derive a new inequality:

$$\begin{aligned}&- \frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &\quad - \frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}_i^{(t)} - \mathbf{w}^*, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &\leq \frac{\eta_t}{m}\sum_{i\in\mathcal{S}(t)} (\frac{1}{\eta_t}\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 + \eta_t\|\nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2) \\ &\quad - \frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}_i^{(t)} - \mathbf{w}^*, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > \\ &= \frac{1}{m}\sum_{i\in\mathcal{S}(t)} \|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 + \frac{\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)} \|\nabla F_i(\mathbf{w}_i^{(t)}, D_i^*)\|^2 \\ &\quad - \frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)} < \mathbf{w}_i^{(t)} - \mathbf{w}^*, \nabla F_i(\mathbf{w}_i^{(t)}, D_i^*) > . \quad (33)\end{aligned}$$

Due to the Lemma A.1,we can get a new inequality:

$$\begin{aligned}
&\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}^{(t)}-\mathbf{w}_i^{(t)}\|^2+\frac{\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}\|\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)\|^2\\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}<\mathbf{w}_i^{(t)}-\mathbf{w}^*,\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)>\\
\leq&\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}^{(t)}-\mathbf{w}_i^{(t)}\|^2\\
&+\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)\\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}<\mathbf{w}_i^{(t)}-\mathbf{w}^*,\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)>.
\end{aligned}\tag{34}$$

Due to the μ-convexity of F_i (Assumption 4.2):

$$\begin{aligned}
&\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}^{(t)}-\mathbf{w}_i^{(t)}\|^2\\
&+\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)\\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}<\mathbf{w}_i^{(t)}-\mathbf{w}^*,\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)>\\
\leq&\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}^{(t)}-\mathbf{w}_i^{(t)}\|^2+\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}\\
&(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}\\
&[(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))+\frac{\mu}{2}\|\mathbf{w}_i^{(t)}-\mathbf{w}^*\|^2].
\end{aligned}\tag{35}$$

Due to the Lemma A.2:

$$\begin{aligned}
&\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}^{(t)}-\mathbf{w}_i^{(t)}\|^2\\
&+\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)\\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}[(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))\\
&+\frac{\mu}{2}\|\mathbf{w}_i^{(t)}-\mathbf{w}^*\|^2]\\
\leq&16\eta_t^2\tau^2G^2-\frac{\eta_t\mu}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^*\|^2\\
&+\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)\\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*)).
\end{aligned}\tag{36}$$

Next, in expectation, $\mathbb{E}[A_2]=0$ due to the unbiased gradient. Then, we obtain the upper bound of A_3 (30).

$$\begin{aligned}
&\eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)\|^2\\
=&\frac{\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}\|\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)\|^2\\
\leq&\frac{2L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*).
\end{aligned}\tag{37}$$

Finally, we can bound A_4 by using the bound of variance of stochastic gradients (Assumption 4.3):

$$\begin{aligned}
&\mathbb{E}[\eta_t^2\|\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)-\overline{\mathbf{g}}^{(t)}\|^2]\\
=&\eta_t^2\mathbb{E}[\|\sum_{i\in\mathcal{S}(t)}\frac{1}{m}(\mathbf{g}_i(\mathbf{w}_i^{(t)},\xi_i^{(t)})-\nabla F_i(\mathbf{w}_i^{(t)},D_i^*))\|^2]\\
=&\frac{\eta_t^2}{m^2}\mathbb{E}_{\mathcal{S}(t)}[\sum_{i\in\mathcal{S}(t)}\mathbb{E}\|\mathbf{g}_i(\mathbf{w}_i^{(t)},\xi_i^{(t)})-\nabla F_i(\mathbf{w}_i^{(t)},D_i^*)\|^2]\\
\leq&\frac{\eta_t^2\sigma^2}{m}.
\end{aligned}\tag{38}$$

Using the upper bounds of A_1 (36), A_2, A_3 (37) and A_4 (38), we have that the expectation of equation (27) is bounded as:

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2]\\
\leq&\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]-\frac{\eta_t\mu}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^*\|^2]\\
&+16\eta_t^2\tau^2G^2+\frac{\eta_t^2\sigma^2}{m}\\
&+\frac{4L\eta_t^2}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)]\\
&-\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))].
\end{aligned}\tag{39}$$

Due to the Lemma A.3, we can get the bound of equation

(39):

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2] \\
\leq &(1-\eta_t\mu)\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]+16\eta_t^2\tau^2G^2+\frac{\eta_t^2\sigma^2}{m} \\
&+\frac{4L\eta_t^2}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)] \\
&-\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))].
\end{aligned} \tag{40}$$

Then we defined:

$$\begin{aligned}
A_5=&\frac{4L\eta_t^2}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)] \\
&-\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))].
\end{aligned} \tag{41}$$

We can represent A_5 in a different form as:

$$\begin{aligned}
&\mathbb{E}[\frac{4L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*) \\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))] \\
=&\mathbb{E}[\frac{4L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}F_i(\mathbf{w}_i^{(t)},D_i^*)-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}F_i(\mathbf{w}_i^{(t)},D_i^*) \\
&-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i^*-F_i(\mathbf{w}^*,D_i^*))+\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}F_i^* \\
&-\frac{4L\eta_t^2}{m}\sum_{i\in\mathcal{S}(t)}F_i^*] \\
=&\mathbb{E}[\frac{2\eta_t(2L\eta_t-1)}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)] \\
&+2\eta_t\mathbb{E}[\frac{1}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^*,D_i^*)-F_i^*)].
\end{aligned} \tag{42}$$

Now, we defined:

$$A_6=\mathbb{E}[\frac{2\eta_t(2L\eta_t-1)}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)]. \tag{43}$$

Now, with $\eta_t<\frac{1}{4L}$ and $v_t=2\eta_t(1-2L\eta_t)$, A_6 can be bounded as:

$$\begin{aligned}
&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^{(t)},D_i^*) \\
&+F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*) \\
=&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^{(t)},D_i^*)) \\
&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*).
\end{aligned} \tag{44}$$

Due to the μ-convexity (Assumption 4.2):

$$\begin{aligned}
&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^{(t)},D_i^*)) \\
&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*) \\
\leq&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}\Big[<\nabla F_i(\mathbf{w}^{(t)},D_i^*),\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}> \\
&+\frac{\mu}{2}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2\Big]-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*).
\end{aligned} \tag{45}$$

Due to the Lemma A.1 and the AM-GM inequality and Cauchy–Schwarz inequality:

$$\begin{aligned}
&-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}\Big[<\nabla F_i(\mathbf{w}^{(t)},D_i^*),\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}> \\
&+\frac{\mu}{2}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2\Big]-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*) \\
\leq&\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}\Big[\eta_tL(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)+(\frac{1}{2\eta_t}-\frac{\mu}{2}) \\
&\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2\Big]-\frac{v_t}{m}\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*) \\
=&-\frac{v_t}{m}(1-\eta_tL)\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*) \\
&+(\frac{v_t}{2\eta_tm}-\frac{v_t\mu}{2m})\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2.
\end{aligned} \tag{46}$$

We can easily prove that $\frac{v_t(1-\eta_t\mu)}{2\eta_t}\leq 1$, thus we can

obtain:

$$
\begin{aligned}
&-\frac{v_t}{m}(1-\eta_t L)\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)\\
&+(\frac{v_t}{2\eta_t m}-\frac{v_t\mu}{2m})\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2\\
\leq&-\frac{v_t}{m}(1-\eta_t L)\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)\\
&+\frac{1}{m}\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2.
\end{aligned}
\tag{47}
$$

By using equation (47), we can bound A5 (41) as:

$$
\begin{aligned}
&\frac{4L\eta_t^2}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)]\\
&-\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))]\\
\leq&\frac{1}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}\|\mathbf{w}_i^{(t)}-\mathbf{w}^{(t)}\|^2]-\frac{v_t}{m}(1-\eta_t L)\\
&\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)]+\frac{2\eta_t}{m}\\
&\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^*,D_i^*)-F_i^*)]\\
\leq&16\eta_t^2\tau^2G^2-\frac{v_t}{m}(1-\eta_t L)\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)]\\
&+\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^*,D_i^*)-F_i^*)].
\end{aligned}
\tag{48}
$$

Due to the Definition 4.1 and 4.2, we can get:

$$
\begin{aligned}
&16\eta_t^2\tau^2G^2-\frac{v_t}{m}(1-\eta_t L)\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^{(t)},D_i^*)-F_i^*)]\\
&+\frac{2\eta_t}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}^*,D_i^*)-F_i^*)]\\
=&16\eta_t^2\tau^2G^2-v_t(1-\eta_t L)\mathbb{E}[\rho(\mathcal{S}(t),\mathbf{w}^{(t)})(F(\mathbf{w}^{(t)},D_i^*)\\
&-\sum_{i=1}^N\mathcal{P}_iF_i^*)]+2\eta_t\mathbb{E}[\rho(\mathcal{S}(t),\mathbf{w}^*)(F^*-\sum_{i=1}^N\mathcal{P}_iF_i^*)]\\
\leq&16\eta_t^2\tau^2G^2-v_t(1-\eta_t L)\overline{\rho}(\mathbb{E}[F(\mathbf{w}^{(t)},D_i^*)]-\sum_{i=1}^N\mathcal{P}_iF_i^*)\\
&+2\eta_t\widetilde{\rho}\Gamma.
\end{aligned}
\tag{49}
$$

Then, we define:

$$
A_7=-v_t(1-\eta_t L)\overline{\rho}(\mathbb{E}[F(\mathbf{w}^{(t)},D_i^*)]-\sum_{i=1}^N\mathcal{P}_iF_i^*).
\tag{50}
$$

We can expand A_7 (50) as:

$$
\begin{aligned}
A_7=&-v_t(1-\eta_t L)\overline{\rho}(\mathbb{E}[F(\mathbf{w}^{(t)},D_i^*)]-\sum_{i=1}^N\mathcal{P}_iF_i^*)\\
=&-v_t(1-\eta_t L)\overline{\rho}\sum_{i=1}^N\mathcal{P}_i(\mathbb{E}[F_i(\mathbf{w}^{(t)},D_i^*)]\\
&-F^*+F^*-F_i^*).
\end{aligned}
\tag{51}
$$

$$
\begin{aligned}
A_7=&-v_t(1-\eta_t L)\overline{\rho}\sum_{i=1}^N\mathcal{P}_i(\mathbb{E}[F_i(\mathbf{w}^{(t)},D_i^*)]-F^*)\\
&-v_t(1-\eta_t L)\overline{\rho}\sum_{i=1}^N\mathcal{P}_i(F^*-F_i^*)\\
=&-v_t(1-\eta_t L)\overline{\rho}(\mathbb{E}[F(\mathbf{w}^{(t)},D^*)]-F^*)\\
&-v_t(1-\eta_t L)\overline{\rho}\Gamma.
\end{aligned}
\tag{52}
$$

Due to the μ-convexity (Assumption 4.2), we can get:

$$
A_7\leq\frac{-v_t(1-\eta_t L)\mu\overline{\rho}}{2}\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]-v_t(1-\eta_t L)\overline{\rho}\Gamma.
\tag{53}
$$

We can easily prove that $-2\eta_t(1-2L\eta_t)(1-\eta_t L)\leq-\frac{3}{4}\eta_t$ and $-(1-2L\eta_t)(1-\eta_t L)\leq-(1-3L\eta_t)$, thus we can obtain:

$$
\begin{aligned}
A_7\leq&-\frac{3\eta_t\mu\overline{\rho}}{8}\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]-2\eta_t(1-2L\eta_t)\\
&(1-\eta_t L)\overline{\rho}\Gamma.\\
\leq&-\frac{3\eta_t\mu\overline{\rho}}{8}\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]-2\eta_t\overline{\rho}\Gamma+6\eta_t^2\overline{\rho}L\Gamma
\end{aligned}
\tag{54}
$$

Thus we can bound A_5 as:

$$
\begin{aligned}
&\frac{4L\eta_t^2}{m}\mathbb{E}[\sum_{i\in\mathcal{S}(t)}(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i^*)-\frac{2\eta_t}{m}\sum_{i\in\mathcal{S}(t)}\\
&(F_i(\mathbf{w}_i^{(t)},D_i^*)-F_i(\mathbf{w}^*,D_i^*))]\\
\leq&-\frac{3\eta_t\mu\overline{\rho}}{8}\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]+2\eta_t\Gamma(\widetilde{\rho}-\overline{\rho})\\
&+\eta_t^2(6\overline{\rho}L\Gamma+16\tau^2G^2).
\end{aligned}
\tag{55}
$$

Now, we can bound $\mathbb{E}[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2]$ as:

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|^2]\leq&\left[1-\eta_t\mu(1+\frac{3\overline{\rho}}{8})\right]\\
&\mathbb{E}[\|\mathbf{w}^{(t)}-\mathbf{w}^*\|^2]\\
&+\eta_t^2(32\tau^2G^2+\frac{\sigma^2}{m}+6\overline{\rho}L\Gamma)\\
&+2\eta_t\Gamma(\widetilde{\rho}-\overline{\rho}).
\end{aligned}
\tag{56}
$$

By defining $\Delta_{t+1} = \mathbb{E}[\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2]$, $B = (1 + \frac{3\overline{\rho}}{8})$, $C = 32\tau^2 G^2 + \frac{\sigma^2}{m} + 6\overline{\rho} L \Gamma$, $D = 2\Gamma(\widetilde{\rho} - \overline{\rho})$, we have that:

$$\Delta_{t+1} \leq (1 - \eta_t \mu B)\Delta_t + \eta_t^2 C + \eta_t D. \tag{57}$$

By setting $\Delta_t \leq \frac{\psi}{t+\gamma}$, $\eta_t = \frac{\beta}{t+\gamma}$, $\beta > \frac{1}{\mu B}$ and $\gamma > 0$, we have that by induction:

$$\begin{aligned}\psi =& \max\left\{\gamma\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2, \right. \\ & \left. \frac{1}{\beta\mu B - 1}(\beta^2 C + D\beta(t+\gamma))\right\}. \end{aligned} \tag{58}$$

Then by the L-smoothness of $F(\cdot)$ (Assumption 4.1), we can get that:

$$\mathbb{E}[F(\mathbf{w}^{(t)}, D^*)] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{\psi}{\gamma + t}. \tag{59}$$

C. Comparison of Dataset Pruning Methods on Data Distribution Heterogeneity

Notation: Necessary notations are introduced as follows.

- $\nabla F_i(\mathbf{w}, D_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \nabla \ell\left(f_{\mathbf{w}}(\mathbf{x}_{ij}), y_{ij}\right)$.
- n_i^{re}: number of samples on client c_i after pruning.
- n^{re}: number of samples on all clients after pruning.
- n_k^{re}: number of samples belonging to target k on all clients after pruning.

According to Definition 4.1, We can obtain a norm Γ to measure the heterogeneity of data distribution. A smaller gamma represents lower heterogeneity in data distribution. Now, we introduce a assumptions utilized for our analysis.

Assumption C.1. *Assuming the total number of clients is sufficiently large, due to the law of large numbers and the consistency of the pruning strategy, the distribution of the global dataset formed across all clients remains nearly unchanged before and after pruning, even if data distribution on each clients is unbalanced, i.e.* $\frac{n_k^{\text{re}}}{n^{\text{re}}} = \frac{n^k}{n}$

In FL scenarios, a large number of clients are typically involved, making our Assumption C.1 reasonable.

Assumption C.2. *When the number of samples on the client is sufficiently large, the data distribution remains almost unchanged before and after pruning. i.e.* $\frac{n_{ik}^{\text{re}}}{n_i^{\text{re}}} = \frac{n_k^{\text{re}}}{n^{\text{re}}}$

According to the law of large numbers, it can be proven that Assumption C.2 is reasonable.

When t is sufficiently large, the weight of the global model is:

$$\begin{aligned}\mathbf{w}^{(t+1)} &= \mathbf{w}^{(t)} - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{w}_i^{(t)}, D_i) \\ &= \mathbf{w}^{(t)} - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}}\frac{1}{n_i}\sum_{j=1}^{n_i}\nabla\ell\left(f_{\mathbf{w}^{(t)}}(x_{ij}), y_{ij}\right) \\ &= \mathbf{w}^{(t)} - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}}\frac{1}{n_i}\sum_{k\in[\mathcal{K}]} \\ &\quad \sum_{j=1}^{n_i}\mathbb{I}(y_{ij} = k)\nabla\ell\left(f_{\mathbf{w}^{(t)}}(x_{ij}), y_{ij}\right). \end{aligned} \tag{60}$$

Then, we defined:

$$\overline{\nabla\ell_i^k} = \frac{1}{n_i^k}\sum_{j=1}^{n_i}\mathbb{I}(y_{ij} = k)\nabla\ell\left(f_{\mathbf{w}^{(t)}}(x_{ij}), y_{ij}\right). \tag{61}$$

Substituting Equation (61) into Equation (60), we get:

$$\begin{aligned}\mathbf{w}^{(t+1)} =& \mathbf{w}^{(t)} - \frac{\eta_t}{m}\sum_{i\in\mathcal{S}}\frac{1}{n_i}\sum_{k\in[\mathcal{K}]} n_i^k\overline{\nabla\ell_i^k} \\ =& \mathbf{w}^{(t)} - \eta_t\sum_{k\in[\mathcal{K}]}\sum_{i\in\mathcal{S}}\frac{n_i^k}{mn_i}\overline{\nabla\ell_i^k}. \end{aligned} \tag{62}$$

Then, the expectation of $\mathbf{w}^{(t+1)}$ is:

$$\mathbb{E}[\mathbf{w}^{(t+1)}] = \mathbb{E}[\mathbf{w}^{(t)} - \eta_t\sum_{k\in[\mathcal{K}]}\sum_{i\in\mathcal{S}}\frac{n_i^k}{mn_i}\overline{\nabla\ell_i^k}]. \tag{63}$$

Then, we define:

$$\overline{\nabla\ell^k} = \sum_{i\in\mathcal{S}}\frac{n_i^k}{n^k}\overline{\nabla\ell_i^k}, \tag{64}$$

where $n^k = \sum_{i\in\mathcal{S}} n_i^k$, then substituting Equation (63) into Equation (64):

$$\begin{aligned}\mathbb{E}[\mathbf{w}^{(t+1)}] =& \mathbb{E}[\mathbf{w}^{(t)} - \eta_t\sum_{k\in\mathcal{K}}\frac{n^k}{mn_i}\overline{\nabla\ell^k}] \\ =& \mathbf{w}^{(t)} - \eta_t\mathbb{E}[\sum_{k\in\mathcal{K}}\frac{n^k}{mn_i}\overline{\nabla\ell^k}] \\ =& \mathbf{w}^{(t)} - \eta_t\frac{1}{\mathcal{K}}\sum_{k=1}^{\mathcal{K}}\mathbb{E}[\overline{\nabla\ell^k}]. \end{aligned} \tag{65}$$

The weight of local model are:

$$\begin{aligned}\mathbf{w}_i^{(t+1)} =&\mathbf{w}_i^{(t)} - \eta_t \nabla F_i(\mathbf{w}_i^t, D_i)\\ =&\mathbf{w}_i^{(t)} - \eta_t \sum_{j=1}^{n_i} \nabla \ell \left(f_{\mathbf{w}^{(t)}}\left(x_{ij}\right), y_{ij}\right)\\ =&\mathbf{w}_i^{(t)} - \eta_t \sum_{k=1}^{\mathcal{K}} \frac{n_i^k}{n_i} \overline{\nabla \ell_i^k}. \end{aligned} \tag{66}$$

C.1. Effective of Double Pruning on Data Distribution Heterogeneity

For dataset pruning methods that can select representative samples from each class, we can get $\mathbb{E}[\nabla \ell^k] \approx \overline{\nabla \ell_i^k}$. According to Assumption C.1, the closer $\frac{n_i^k}{n_i}$ to $\frac{1}{\mathcal{K}}$, the closer $\mathbf{w}_i^{(t+1)}$ to $\mathbf{w}^{(t+1)}$.

Double Pruning significantly reduces samples in the "Large-capacity Class," bringing $\mathbf{w}_i^{(t+1)}$ closer to $\mathbf{w}^{(t+1)}$, which ultimately brings $F_i(D_i^*, \mathbf{w}^*)$ closer to $F_i(D_i^*, \mathbf{w}_i^*)$ and reduce Γ (Equation (13)).

C.2. Effective of Different Coreset Selection Methods on Data Distribution Heterogeneity

We can divide n_i^k samples of class k on client i into $\mathcal{N}_i^r$ redundant samples, $\mathcal{N}_i^d$ decision boundary samples, and $\mathcal{N}_i^h$ hard samples. Then the average gradient of class k on client i can be represented as :

$$\overline{\nabla \ell_i^k} = \frac{1}{n_i^k}[\mathcal{N}_i^r \overline{\nabla \ell_i^k}^r + \mathcal{N}_i^d \overline{\nabla \ell_i^k}^d + \mathcal{N}_i^h \overline{\nabla \ell_i^k}^h], \tag{67}$$

where $\mathcal{N}_i^r$, $\overline{\nabla \ell_i^k}^r$, $\mathcal{N}_i^d$, $\overline{\nabla \ell_i^k}^d$, $\mathcal{N}_i^h$, $\overline{\nabla \ell_i^k}^h$ represent the number and the average gradients produced by the redundant samples, decision boundary samples, and hard samples, respectively, for class k on client i. Global average gradient of class k can be represented as :

$$\mathbb{E}[\overline{\nabla \ell^k}] = \frac{1}{n^k}[\mathcal{N}_r \mathbb{E}[\overline{\nabla \ell^k}^r] + \mathcal{N}_d \mathbb{E}[\overline{\nabla \ell^k}^d] + \mathcal{N}_h \mathbb{E}[\overline{\nabla \ell^k}^h], \tag{68}$$

where $\mathcal{N}_r$, $\mathbb{E}[\overline{\nabla \ell^k}^r]$, $\mathcal{N}_d$, $\mathbb{E}[\overline{\nabla \ell^k}^d]$, $\mathcal{N}_h$, $\mathbb{E}[\overline{\nabla \ell^k}^h]$ represent the global number and the expectation of global average gradients produced by the redundant samples, decision boundary samples, and hard samples, respectively, for class k.

Since redundant samples provide almost no gradient during training, $\mathbb{E}[\|\overline{\nabla \ell^k}\|] \approx \|\overline{\nabla \ell_i^k}\| \approx 0$. As the hard samples contain a large number of extremely rare examples, the unique extreme information in these samples often causes the resulting gradients to be noisy, we rewrite average gradient produced by the hard samples as:

$$\mathcal{N}_i^h \overline{\nabla \ell_i^k}^h = \mathcal{N}_i^h \overline{\nabla \ell_i^k}^{h*} + \overline{\nabla \ell_i^k}^n, \tag{69}$$

Table 5. The Top-1 accuracy (%) of coreset selection methods with VGG-11 on CINIC-10.

Methods	$\alpha = 0.1, p_i^l = 0.1$			
	$p_i^f = 0.3$	$p_i^f = 0.5$	$p_i^f = 0.7$	$p_i^f = 0.9$
EL2N	56.23±0.96	53.51±1.08	50.23±1.42	46.76±1.45
Moderate	56.76±0.72	55.76±0.79	55.16±0.82	52.73±0.87
GM	57.03±0.60	56.74±0.63	55.68±0.71	54.02±0.78
GradND	55.80±1.18	53.02±1.24	49.68±1.32	47.18±1.37
Forgetting	53.49±0.80	51.93±1.09	50.67±1.40	49.81±1.56
Random	56.65±1.02	55.78±1.61	54.46±2.06	53.07±2.34
FedCS(ours)	**57.58**±0.83	**57.39**±0.91	**56.72**±0.95	**55.08**±0.97
Whole Dataset	57.03±0.61	57.03±0.61	57.03±0.61	57.03±0.61

$$\mathcal{N}_h \mathbb{E}[\overline{\nabla \ell^k}^h] = \mathcal{N}_h \mathbb{E}[\overline{\nabla \ell^k}^{h*}] + \mathbb{E}[\overline{\nabla \ell^k}^n]. \tag{70}$$

If DC Score is used to select samples near the decision boundary, the weight of local model is:

$$\mathbf{w}_i^{(t+1)} \approx \mathbf{w}_i^{(t)} - \eta_t \sum_{k=1}^{\mathcal{K}} \frac{n_{ik}^{\mathrm{re}}}{n_i^{\mathrm{re}}} \overline{\nabla \ell_i^k}^d, \tag{71}$$

where n_{ik}^{re} denotes the num of samples belongs to targert k on client c_i after pruning.

The weights of global model is:

$$\mathbf{w}^{(t+1)} \approx \mathbf{w}^{(t)} - \eta_t \sum_{k=1}^{\mathcal{K}} \frac{n_k^{\mathrm{re}}}{n^{\mathrm{re}}} \mathbb{E}[\overline{\nabla \ell^k}^d], \tag{72}$$

$$\mathbb{E}[\overline{\nabla \ell^k}^d] = \mathbb{E}[\sum_{i \in \mathcal{S}} \overline{\nabla \ell_i^k}^d]. \tag{73}$$

When the number of samples is sufficiently large, $E[\overline{\nabla \ell^k}^d]$ and $\overline{\nabla \ell_i^k}^d$ will be very close.

If a strategy for selecting hard samples is used for core set selection, the weight of local model is:

$$\mathbf{w}_i^{(t+1)} \approx \mathbf{w}_i^{(t)} - \frac{\eta_t}{n_i^{\mathrm{re}}} \sum_{k=1}^{\mathcal{K}} (n_{ik}^{\mathrm{re}} \overline{\nabla \ell_i^k}^h + \overline{\nabla \ell_i^k}^n). \tag{74}$$

The weight of global model is:

$$\mathbf{w}^{(t+1)} \approx \mathbf{w}^{(t)} - \frac{\eta_t}{n^{\mathrm{re}}} \sum_{k=1}^{\mathcal{K}} (n_k^{\mathrm{re}} \mathbb{E}[\overline{\nabla \ell^k}^h] + \mathbb{E}[\overline{\nabla \ell^k}^n]). \tag{75}$$

Since noise gradients are generated randomly, the direction of $\mathbb{E}[\overline{\nabla \ell^k}^n]$ and $\overline{\nabla \ell_i^k}^n$ will be significantly different.

Therefore, compared to the pruning strategy that selects hard samples, the differences between $\frac{n_{ik}^{\mathrm{re}}}{n_i^{\mathrm{re}}}$ and $\frac{n_k^{\mathrm{re}}}{n^{\mathrm{re}}}$ are almost identical, but the difference between $\mathbf{w}_i^{(t+1)}$ and $\mathbf{w}_i^{(t)}$ in DC Score is smaller, resulting in more similar $\mathbf{w}_i^{(t+1)}$ and $\mathbf{w}^{(t+1)}$ ultimately leading to a smaller Γ (Equation (13)).

D. Supplementary Experiments

VGG-11 on CINIC-10. CINIC-10 extends CIFAR-10 with the down-sampled ImageNet images consisting of 90K training images and 90K testing images. As shown in Table 5, our proposed method demonstrates the best performance across various pruning rates with an accuracy improvement of 0.55%-8.32% over other methods. Although overfitting does not occur, the accuracy of FedCS with a low pruning rate still exceeds that of training with the entire dataset when all clients participate in update. It further confirms the superiority of FedCS compared to other SOTA methods.