Appendix

A. Proof of Theorem 1

Theorem 1 *(Global Loss Reduction) With Assumption on Lipschitz property, for an arbitrary set of clients $C_t \subseteq C$ selected by the server in round t, the reduction of global loss $F(w)$ is bounded by:*

$$\begin{aligned}\underbrace{F(w_{\text{fed}}^{t-1})-F(w_{\text{fed}}^{t})}_{\text{global loss reduction}} \geqslant \sum_{c\in C_t}\sum_{i=0}^{m-1}\sum_{(x,y)\in B_c}\big[-\alpha_c \underbrace{\| \nabla_w l(w_c^{t,i},x,y)\|_2^2}_{\text{term 1}} \\ + \beta_c \underbrace{\langle \nabla_w l(w_c^{t,i},x,y), \nabla_w F(w_{\text{fed}}^{t-1})\rangle}_{\text{term 2}}\big],\end{aligned} \tag{15}$$

where $\alpha_c = \frac{L\zeta_c^t}{2}\cdot\left(\frac{\eta}{|B_c|}\right)^2$ and $\beta_c = \zeta_c^t \cdot \left(\frac{\eta}{|B_c|}\right)$.

Proof. From the L-Lipschitz continuity of global loss function $F(w)$, we have

$$\begin{aligned}&F(w_{\text{fed}}^{t}) - F(w_{\text{fed}}^{t-1}) \\ &\leqslant \langle \nabla_w F(w_{\text{fed}}^{t-1}), w_{\text{fed}}^{t} - w_{\text{fed}}^{t-1}\rangle + \frac{L}{2}\| w_{\text{fed}}^{t} - w_{\text{fed}}^{t-1}\|^2,\end{aligned} \tag{16}$$

where the global model can be further decomposed into the weighted sum of the updated models of participating clients:

$$\begin{aligned}&(16) \\ &= \langle \nabla_w F(w_{\text{fed}}^{t-1}), \sum_{c\in C_t}\zeta_c^t w_c^{t,m} - w_{\text{fed}}^{t-1}\rangle + \frac{L}{2}\|\sum_{c\in C_t}\zeta_c^t w_c^{t,m} - w_{\text{fed}}^{t-1}\|^2 \\ &\leqslant \sum_{c\in C_t}\zeta_c^t \langle \nabla_w F(w_{\text{fed}}^{t-1}), w_c^{t,m} - w_{\text{fed}}^{t-1}\rangle + \frac{L}{2}\sum_{c\in C_t}\zeta_c^t \| w_c^{t,m} - w_{\text{fed}}^{t-1}\|^2 .\end{aligned} \tag{17}$$

We can further express the locally updated model $w_c^{t,m}$ of each client $c\in C_t$ as the difference between the original global model w_{fed}^{t-1} and m local model updates:

$$\begin{aligned}(17) &\leqslant \sum_{c\in C_t}\zeta_c^t\Big[\langle \nabla_w F(w_{\text{fed}}^{t-1}), -\eta\sum_{i=0}^{m-1}\nabla_w \tilde{F}_c(w_c^{t,i})\rangle \\ &\quad + \frac{L}{2}\|\sum_{i=0}^{m-1}\eta\nabla_w \tilde{F}(w_c^{t,i})\|^2\Big] \\ &= \sum_{c\in C_t}\zeta_c^t\Big[\frac{L}{2}\|\sum_{i=0}^{m-1}\eta\nabla_w \tilde{F}(w_c^{t,i})\|^2 \\ &\quad - \eta\sum_{i=0}^{m-1}\langle \nabla_w F(w_{\text{fed}}^{t-1}), \nabla_w \tilde{F}_c(w_c^{t,i})\rangle\Big],\end{aligned} \tag{18}$$

where the i-th local model update of each client c can be expressed as the average gradient of the stored data samples:

$$\begin{aligned}(18) &= \sum_{c\in C_t}\|\sum_{i=0}^{m-1}\eta\sum_{x,y\in B_c}\frac{1}{|B_c|}\nabla_w l(w_c^{t,i},x,y)\|^2 - \\ &\quad \frac{\eta}{|B_c|}\sum_{i=0}^{m-1}\sum_{(x,y)\in B_c}\langle \nabla_w F(w_{\text{fed}}^{t-1}), \nabla_w l(w_c^{t,i},x,y)\rangle\Big] \\ &\leqslant \sum_{c\in C_t}\Bigg[\sum_{i=0}^{m-1}\sum_{x,y\in B_c}\Big(\alpha_c \|\nabla_w l(w_c^{t,i},x,y)\|^2 \\ &\quad - \beta_c\langle \nabla_w F(w_{\text{fed}}^{t-1}), \nabla_w l(w_c^{t,i},x,y)\rangle\Big)\Bigg].\end{aligned}$$

where $\alpha_c = \frac{L\zeta_c^t}{2}\cdot\left(\frac{\eta}{|B_c|}\right)^2$ and $\beta_c = \zeta_c^t \cdot \left(\frac{\eta}{|B_c|}\right)$. □

B. Proof of Theorem 2

Theorem 2 *(Model Weight Divergence) With Assumption on Lipschitz property, for an arbitrary participating client set C_t, we have the following inequality for the weight divergence between the models trained through FL and CL after the t-th training round.*

$$\begin{aligned}\| w_{\text{fed}}^{t} - w_{\text{cen}}^{mt}\|_2 &\leqslant (1+\eta L)^m \| w_{\text{fed}}^{t-1} - w_{\text{cen}}^{m(t-1)}\|_2 \\ &+ \sum_{c\in C_t}\zeta_c^t\left[\eta\sum_{i=0}^{m-1}(1+\eta L)^{m-1-i}G_c(w_c^{t,i})\right],\end{aligned} \tag{19}$$

where $G_c(w) = \| \nabla_w \tilde{F}_c(w) - \nabla_w F(w)\|_2$.

Proof. According to the aggregation formula of Fed-Avg and the model update formula (1), we have:

$$\begin{aligned}\| w_{\text{fed}}^{t} - w_{\text{cen}}^{mt}\| &= \|\sum_{c\in C_t}\zeta_c^t w_c^{t,m} - w_{\text{cen}}^{mt}\| \\ &= \|\sum_{c\in C_t}\zeta_c^t\left[w_c^{t,m-1} - \eta\nabla_w \tilde{F}_c(w_c^{t,m-1})\right] \\ &\quad - w_{\text{cen}}^{mt-1} + \eta\nabla_w F(w_{\text{cen}}^{mt-1})\| \\ &\leqslant \|\sum_{c\in C_t}\zeta_c^t w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| \\ &\quad + \eta\|\sum_{c\in C_t}\zeta_c^t\nabla\tilde{F}_c(w_c^{t,m-1}) - \nabla F(w_{\text{cen}}^{mt-1})\| .\end{aligned} \tag{20}$$

Then, we leverage the triangle inequality to upper bound (20):

$$\begin{aligned}(20) &\leqslant \sum_{c\in C_t}\zeta_c^t\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| + \eta\|\sum_{c\in C_t}\zeta_c^t\Big[\nabla_w\tilde{F}_c(w_c^{t,m-1}) \\ &\quad - \nabla_w F(w_c^{t,m-1}) + \nabla_w F(w_c^{t,m-1}) - \nabla_w F(w_{\text{cen}}^{mt-1})\Big]\| \\ &= \sum_{c\in C_t}\zeta_c^t\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| \\ &\quad + \eta\|\sum_{c\in C_t}\zeta_c^t\big[\nabla_w\tilde{F}_c(w_c^{t,m-1}) - \nabla_w F(w_c^{t,m-1})\big]\| \\ &\quad + \eta\|\sum_{c\in C_t}\zeta_c^t\big[\nabla_w F(w_c^{t,m-1}) - \nabla_w F(w_{\text{cen}}^{mt-1})\big]\| .\end{aligned} \tag{21}$$

According to the Lipschitz continuity of global model F in Assumption 1, we can further upper bound the Eq. (21):

$$\begin{aligned}(21) &\leqslant \sum_{c\in C_t}\zeta_c^t(1+\eta L)\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| + \\ &\quad \eta\|\sum_{c\in C_t}\zeta_c^t\big[\nabla_w\tilde{F}_c(w_c^{t,m-1}) - \nabla_w F(w_c^{t,m-1})\big]\| \\ &\leqslant \sum_{c\in C_t}\zeta_c^t\Big[(1+\eta L)\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| + \eta G_c(w_c^{t,m-1})\Big],\end{aligned} \tag{22}$$

where $G_c(w) = \| \nabla_w \tilde{F}_c(w) - \nabla_w F(w)\|$ intuitively represents the divergence between empirical losses over the local data distribution and global data distribution.

Next, we derive the upper bound of $\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\|$ for each participating client $c\in C_t$:

$$\begin{aligned}&\| w_c^{t,m-1} - w_{\text{cen}}^{mt-1}\| \\ &= \| w_c^{t,m-2} - \eta\nabla_w\tilde{F}_c(w_c^{t,m-2}) - w_{\text{cen}}^{mt-2} + \eta\nabla_w F(w_{\text{cen}}^{mt-2})\| \\ &\leqslant \| w_c^{t,m-2} - w_{\text{cen}}^{mt-2}\| + \eta\|\nabla_w F(w_c^{t,m-2}) - \\ &\quad \nabla_w F(w_{\text{cen}}^{mt-2})\| + \eta\|\nabla_w\tilde{F}_c(w_c^{t,m-2}) - \nabla_w F(w_c^{t,m-2})\|,\end{aligned} \tag{23}$$

which can be transformed using Assumption of Lipschitz:

$$
\begin{aligned}
(23) \leqslant & (1+\eta L) \parallel w_c^{t,m-2} - w_{\text{cen}}^{mt-2} \parallel + \eta G_c(w_c^{t,m-2}) \\
\leqslant & (1+\eta L)^{m-1} \parallel w_c^{t,0} - w_{\text{cen}}^{(t-1)m} \parallel + \\
& \eta \sum_{i=0}^{m-2} (1+\eta L)^{m-2-i} G_c(w_c^{t,i})
\end{aligned} \tag{24}
$$

Combining inequalities (22) and (24), we have:

$$
\begin{aligned}
\parallel w_{\text{fed}}^{t} - w_{\text{cen}}^{mt} \parallel \leqslant & \sum_{c \in C_t} \zeta_c^t \Big[(1+\eta L)^m \parallel w_c^{t,0} - w_{\text{cen}}^{(t-1)m} \parallel \\
& + \eta \sum_{i=0}^{m-1} (1+\eta L)^{m-1-i} G_c(w_c^{t,i}) \Big] \\
= & (1+\eta L)^m \parallel w_{\text{fed}}^{t-1} - w_{\text{cen}}^{(t-1)m} \parallel \\
& + \sum_{c \in C_t} \zeta_c^t \Big[\eta \sum_{i=0}^{m-1} (1+\eta L)^{m-1-i} G_c(w_c^{t,i}) \Big].
\end{aligned}
$$

Hence, we bound the divergence between models trained through FL and CL by the initial model difference and the additional divergence caused by m local model updates of heterogeneous participating clients. □

C. Proof of Lemma 1

Lemma 1 *(Gradient Divergence) For an arbitrary client $c \in C$, $G_c(w) = \parallel \nabla \tilde{F}_c(w) - \nabla F(w) \parallel_2$ is bounded by:*

$$
G_c(w) \leqslant \Big[\underbrace{\parallel \nabla_w F(w) \parallel_2^2}_{\text{constant}} + \sum_{(x,y) \in B_c} \frac{1}{|B_c|} \Big(\underbrace{\parallel \nabla_w l(w,x,y) \parallel_2^2}_{\text{term 1}} - 2 \underbrace{\big\langle \nabla_w l(w,x,y), \nabla_w F(w) \big\rangle}_{\text{term 2}} \Big) \Big]^{1/2}, \tag{25}
$$

where $\delta =$ is a constant term for all data samples.

Proof. The main idea of the proof is to express the local gradient $\nabla_w \tilde{F}_c(w)$ as the average gradient of the locally stored data samples in B_c:

$$
\begin{aligned}
& G_c(w) = \parallel \nabla_w \tilde{F}_c(w) - \nabla_w F(w) \parallel \\
= & \Big[\parallel \nabla_w F(w) \parallel^2 + \parallel \nabla_w \tilde{F}_c(w) \parallel^2 - 2 \big\langle \nabla_w F(w), \nabla_w \tilde{F}_c(w) \big\rangle \Big]^{1/2} \\
= & \Big[\parallel \nabla_w F(w) \parallel^2 + \parallel \sum_{(x,y) \in B_c} \frac{\nabla_w l(w,x,y)}{|B_c|} \parallel^2 \\
& - 2 \big\langle \nabla_w F(w), \sum_{(x,y) \in B_c} \frac{\nabla_w l(w,x,y)}{|B_c|} \big\rangle \Big]^{1/2} \\
\leqslant & \Big[\parallel \nabla_w F(w) \parallel^2 + \sum_{(x,y) \in B_c} \frac{1}{|B_c|} \Big(\parallel \nabla_w l(w,x,y) \parallel^2 \\
& - 2 \big\langle \nabla_w F(w), \nabla_w l(w,x,y) \big\rangle \Big) \Big]^{1/2},
\end{aligned}
$$

where $\parallel \nabla_w F(w) \parallel^2$ is a constant term for all clients and local data samples. □