Supplementary: Hierarchical Global Asynchronous Federated Learning Across Multi-Center

A. Common Properties

Properties 1 *The F is L-smooth with Lipschitz constant L if*

$$\|\nabla F(\boldsymbol{u}) - \nabla F(\boldsymbol{v})\| \leq L\|\boldsymbol{u}-\boldsymbol{v}\| \quad \forall \boldsymbol{u}, \boldsymbol{v}. \tag{20}$$

$$F(\boldsymbol{v}) \leq F(\boldsymbol{u}) + \langle \nabla F(\boldsymbol{u}), \boldsymbol{v}-\boldsymbol{u}\rangle + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{u}\|^2 \quad \forall \boldsymbol{u}, \boldsymbol{v}. \tag{21}$$

For any vectors v_i, v_j and $i, j \in \mathbb{Z}^+$ and a constant $\Omega > 0$, it has:

$$\|\sum_{i=1}^{m} v_i\|^2 \leq m(\sum_{i=1}^{m} \|v_i\|^2), \tag{22}$$

$$\|v_i + v_j\|^2 \leq (1+\Omega)\|v_i\|^2 + \left(1+\Omega^{-1}\right)\|v_j\|^2. \tag{23}$$

B. Proof of Proposition 1

Proposition 1. For any step t_j within the sub-center update, we denote $\eta_s G_j^{t_j} = w_j^{t_j} - w_j^{t_j-1}$ following the sub-center client model update rule (4). The surrogate learning rate η_s of sub-center satisfies:

$$\eta_s^2 \leq \frac{2+2L^2}{\alpha^2}. \tag{24}$$

Proof Referring from (Acar et al., 2021) Appendix B.1, for the local proximal gradient and the model update in sub-center j, it has the following

$$w_j^{t_j} = \gamma_j^{t_j} - \frac{1}{\alpha}h_j^{t_j}, \quad \gamma_j^{t_j} = \frac{1}{N_j}\sum_{i\in[n_j]} w_{j,i}^{t_j} \quad , \tag{25}$$

where $w_j^{t_j}$ is the sub-model at sub-center fusion step t_j.

And from the property analysis in (Acar et al., 2021) the state $h_j^{t_j}$ becomes the average gradient across clients when $w_j^{t_j}$ converges,

$$h_j^{t_j} = \frac{1}{N_j}\sum_{i\in[n_j]} \nabla F_{j,i}(w_i^{t_j}). \tag{26}$$

Therefore, the model change for sub-center is given by

$$w_j^{t_j+1} - w_j^{t_j} = \gamma_j^{t_j+1} - \gamma_j^{t_j} + \frac{1}{\alpha}(h_j^{t_j} - h_j^{t_j+1}). \tag{27}$$

By utilizing lemma 2 from FedDyn (Acar et al., 2021), we derive the expected value as follow:

$$\mathbb{E}\left[\gamma_j^{t_j} - \gamma_j^{t_j-1}\right] = \frac{1}{\alpha N_j} \sum_{i \in [n_j]} \mathbb{E}\left[-\nabla f_{j,i}(w_{j,i}^{t_j})\right]. \tag{28}$$

Here, we can define the general model change equation as

$$\Delta_j^t = \sum_{q=t_j-R}^{t_j} \eta_s G_j^q = \sum_{q=t_j-R}^{t_j} (\gamma_j^{q+1} - \gamma_j^q + \frac{1}{\alpha}(h_j^q - h_j^{q+1})), \tag{29}$$

where Δ_j^t denote the model parameter accumulated difference of sub-center j by sub-center rounds R at global step t.

Referring from (27) with $\alpha \geq 1$ and inequality (23), we have

$$\mathbb{E}\|w_j^{t_j} - w_j^{t_j-1}\|^2 \leq 2\mathbb{E}\|\gamma_j^{t_j} - \gamma_j^{t_j-1}\|^2 + 2\mathbb{E}\|h_j^{t_j-1} - h_j^{t_j}\|^2, \tag{30}$$

and with L-smooth attribute (20) and (25), we have

$$\mathbb{E}\|h_j^{t_j-1} - h_j^{t_j}\|^2 \leq L^2\mathbb{E}\|\gamma_j^{t_j} - \gamma_j^{t_j-1}\|^2. \tag{31}$$

We define η_s as the sub-center model update step size rate and $G_j^{t_j}$ is the surrogate unbiased estimate gradient of sub-center j. The $\nabla F_j(w_j^{t_j})$ is the gradient from each sub-center j update. From above and (30), we have:

$$\mathbb{E}\|w_j^{t_j} - w_j^{t_j-1}\|^2 = \mathbb{E}\|\eta_s G_j^{t_j}\|^2 = \eta_s^2\mathbb{E}\|\nabla F_j(w_j^{t_j})\|^2 \tag{32}$$

$$\leq \frac{2}{\alpha^2 N_j^2}\| \sum_{i \in [n_j]} \mathbb{E}\left[-\nabla F_{j,i}(w_{j,i}^{t_j})\right] \|^2 + 2\mathbb{E}\|h_j^{t_j-1} - h_j^{t_j}\|^2. \tag{33}$$

Combining (33), (30), (28) and (31), we have

$$\mathbb{E}\|w_j^{t_j} - w_j^{t_j-1}\|^2 \leq 2\mathbb{E}\|\gamma_j^{t_j} - \gamma_j^{t_j-1}\|^2 + 2L^2\mathbb{E}\|\gamma_j^{t_j} - \gamma_j^{t_j-1}\|^2, \tag{34}$$

Following above (34) with (28) and (32), we have

$$\mathbb{E}\|\eta_s G_j^{t_j}\|^2 \leq \frac{2+2L^2}{\alpha^2}\|\frac{1}{N_j} \sum_{i \in [N_j]} \mathbb{E}\left[\nabla F_{j,i}(w_{j,i}^{t_j})\right] \|^2 \tag{35}$$

Therefore, we have following bound for Proposition 1

$$\eta_s^2 \leq \frac{2+2L^2}{\alpha^2}. \tag{36}$$

■

We apply η_s^2 to all sub-centers for the theoretical proof.

C. Properties for Global Convergence

C.1. Bounds on buffered aggregation

In buffered model aggregations, only a subset of sub-center models with buffer size K participate in the model fusion process. We have $w^{t+1} = w^t - \eta_g \frac{1}{K}\Delta^t + \eta_g v^t$, $v^t = \frac{1}{K}\sum_{j\in[H_t]}(c^t - \Delta_j^{t-\Gamma_j^t})$, $c^t = \frac{1}{M}\sum_{j=1}^{M} c_j^t$ and $c_j^t = \Delta_j^{t-\zeta_j^t}$ from Algorithm 1.

Lemma 1 *For any subset $K \subseteq m$ in the global buffer with $|H_t| = K$ and $|m| = M$, the following upper bound holds for the assumption of diversity across sub-center:*

$$\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}[\|\nabla F_j(w) - \nabla F(w)\|^2] \leq \frac{M}{K}\sigma_g^2. \tag{37}$$

Proof
Referring to Assumption 7, we have

$$\begin{aligned}&\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}[\|\nabla F_j(w) - \nabla F(w)\|^2]\\ &\leq \frac{M}{K}\cdot\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}[\|\nabla F_j(w) - \nabla F(w)\|^2] \leq \frac{M}{K}\sigma_g^2.\end{aligned}$$

■

According to the Lemma 3 of FedAdam (Reddi et al., 2021), with $r \in \{0,\dots,R-1\}$ and Lemma 1, the expectation difference between lower level (sub-center) model $w_j^{t,r}$ from global buffer and global w^t after R internal rounds of sub-center, can be reformed to following:

$$\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}[\|w_j^{t-\Gamma_j^t,r} - \boldsymbol{w}^{t-\Gamma_j^t}|^2] \leq 5R\eta_s^2(\sigma_s^2 + 6R\cdot\frac{M}{K}\sigma_g^2) + 30R^2\eta_s^2\mathbb{E}\|\nabla F(w^{t-\Gamma_j^t})\|^2, \tag{38}$$

with sub-center step rate $\eta_s \leq \frac{1}{8RL}$ to satisfy global buffer aggregation and follow Assumption 6 and 7.

Given same conditions as above, for all sub-center model in the state variable c_j:

$$\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}[\|w_j^{t-\zeta_j^t,r} - \boldsymbol{w}^{t-\zeta_j^t}|^2] \leq 5R\eta_s^2(\sigma_s^2 + 6R\sigma_g^2) + 30R^2\eta_s^2\mathbb{E}\|\nabla F(w^{t-\zeta_j^t})\|^2. \tag{39}$$

Extending from (Wang et al., 2023) Lemma G.1 and Inequality of G.16 following (Reddi et al., 2021) and Lemma 1, We derive the bound for the summation of gradients of a subset of sub-center models from global buffer:

$$\begin{aligned}\sum_{j=1}^{K}\|\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\Gamma_j^t,r})\|^2 &\leq 15KR^3L^3\eta_s^2\left(\sigma_s^2 + 6R\frac{M}{K}\sigma_g^2\right)\\ &\quad + \left(90KR^4L^2\eta_s^2\ + 3KR^2\right)\left\|\nabla F(w^{t-\Gamma_j^t})\right\|^2 + 3MR^2\sigma_g^2.\end{aligned} \tag{40}$$

Similar to (39), we can also reformulate (40) using ζ and obtain the bound of summation gradient from state variable:

$$\begin{aligned}\sum_{j=1}^{M}\|\sum_{r=0}^{R-1}\nabla F_i(w_j^{t-\zeta_j^t,r})\|^2 \le 15MR^3L^3\eta_s^2\left(\sigma_s^2+6R\sigma_g^2\right)+\big(90MR^4L^2\eta_s^2\\ +3MR^2\big)\left\|\nabla F(w^{t-\zeta_j^t})\right\|^2+3MR^2\sigma_g^2.\end{aligned} \tag{41}$$

C.2. Bounds for stale model

From (Wang et al., 2023) Equation E.5, it gives a expected bound for current model at step t and stable model at step $t-\tau_i^t$:

$$\mathbb{E}\left[\left\|\boldsymbol{w}^t-\boldsymbol{w}^{t-\tau_i^t}\right\|^2\right]=\mathbb{E}\left\|\sum_{q=t-\tau_i^t}^{t-1}\left(\boldsymbol{w}^{q+1}-\boldsymbol{w}^q\right)\right\|^2\le\tau_{\max}\sum_{q=t-\tau_i^t}^{t-1}\mathbb{E}\left[\left\|\boldsymbol{w}^{q+1}-\boldsymbol{w}^q\right\|^2\right], \tag{42}$$

where τ_i^t denotes the delayed model i at time step t. In our case, we substitute τ with Γ and ζ which represent the model delays in the global buffer and state variables. And the w can be either client worker model or sub-center fusion model.

C.3. Sub-center stale model bounded properties

With the diversity assumption of the sub-center and global gradient

$$\mathbb{E}\|\nabla F_j(w^t)-\nabla F(w^t)\|\le\sigma_g^2. \tag{43}$$

From the smoothness property from (20), we have $\|\nabla F_j(w^{t-\Gamma_j^t})-\nabla F_j(w_j^{t-\Gamma_j^t,r})\|\le L\|w^{t-\Gamma_j^t}-w_j^{t-\Gamma_j^t,r}\|$ and $\|\nabla F_j(w^{t-\zeta_j^t})-\nabla F_j(w_j^{t-\zeta_j^t,r})\|\le L\|w^{t-\zeta_j^t}-w_j^{t-\zeta_j^t,r}\|$, which can be expanded by (38) and (39).

Also, from Wang et al. (2023) Equation E.5 with (42), we have bound the difference between sub-center model $w^{t-\Gamma_j^t}$ in the buffer and global model, and also the bound between state model $w^{t-\zeta_j^t}$ and global model at time step $t-\Gamma_j^t$ and $t-\zeta_j^t$ respectively.

$$\mathbb{E}\|w^t-w^{t-\Gamma_j^t}\|^2=\mathbb{E}\|\sum_{q=t-\Gamma_j^t}^{t-1}(w^{q+1}-w^q)\|^2\le\Gamma_{max}\sum_{q=t-\Gamma_j^t}^{t-1}\mathbb{E}\|w^{q+1}-w^q\|^2, \tag{44}$$

$$\mathbb{E}\|w^t-w^{t-\zeta_j^t}\|^2=\mathbb{E}\|\sum_{q=t-\zeta_j^t}^{t-1}(w^{q+1}-w^q)\|^2\le\zeta_{max}\sum_{q=t-\zeta_j^t}^{t-1}\mathbb{E}\|w^{q+1}-w^q\|^2. \tag{45}$$

D. Global Convergence Analysis

For our HGA-FL algorithm 1, we have following inequality from the smoothness (21)

$$
\begin{aligned}
&\mathbb{E}[F(w^{t+1})] - F(w^t) \\
&\leq \mathbb{E}\left\langle \nabla F(w^t), w^{t+1} - w^t \right\rangle + \frac{L}{2}\mathbb{E}\|w^{t+1} - w^t\|^2 && (46) \\
&= \underbrace{-\eta_g \mathbb{E}[\left\langle \nabla F(w^t), \frac{1}{K}\Delta^t - v^t \right\rangle]}_{\text{T1}} + \underbrace{\frac{\eta_g^2 L}{2}\mathbb{E}\|\Delta^t - v^t\|^2}_{\text{T2}}, && (47)
\end{aligned}
$$

where $\Delta^t = \sum_{j\in[H_t]} \Delta_j^{t-\Gamma_j^t} = \sum_{j\in[H_t]} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t,r})$. Detail variable assigments check Section 2.3.

D.1. Expanding for T_1

For polarization identities vector inner produce equation $\langle a, b\rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2]$ with Equation (29), at each global fusion time step t, the T_1 can extend to be

$$
\begin{aligned}
T_1 &= -\eta_g \mathbb{E}\left[\left\langle \nabla F(w^t)\ ,\ \frac{2}{K}\sum_{j\in[H_t]}\sum_{q=t-\Gamma_j^t}^{t-1} \eta_s G_j^q - \frac{1}{M}\sum_{j=1}^{M}\sum_{q=t-\zeta_j^t}^{t-1} \eta_s G_j^q \right\rangle\right] \\
&= -\eta_g \eta_s R \mathbb{E}\left[\left\langle \nabla F(w^t)\ ,\ \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t,r}) - \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\zeta_j^t,r}) \right\rangle\right].
\end{aligned}
\tag{48}
$$

Thus, the T_1 can be expanded to

$$
\begin{aligned}
T_1 &= -\frac{\eta_g\eta_s R}{2}\mathbb{E}\|\nabla F(w^t)\|^2 && (49) \\
&\quad \underbrace{- \frac{\eta_g\eta_s R}{2}\mathbb{E}\|\frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\Gamma_j^t,r}) - \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\zeta_j^t,r})\|^2}_{T_{1.2}} && (50) \\
&\quad \underbrace{+ \frac{\eta_g\eta_s R}{2}\mathbb{E}\|\nabla F(w^t) - \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\Gamma_j^t,r}) + \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\zeta_j^t,r})\|^2}_{T_{1.3}}. && (51)
\end{aligned}
$$

D.1.1. Expanding the terms in T_1

We define the common terms in T_1 and T_2

$$
V^t = \frac{K+1}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\Gamma_j^t,r}) - \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\zeta_j^t,r}). \tag{52}
$$

With above bound of $\sum_{i=1}^{M_j} \| \sum_{k=0}^{R-1} \nabla F_j(w_j^{t,r}) \|^2$, the $T_{1.2}$ can be

$$T_{1.2} = -\frac{\eta_g \eta_s R}{2} \mathbb{E}\|V^t\|^2. \tag{53}$$

D.2. Expanding for T_2

We observe that $T_{1.2}$ and T_2 have the same term, and

$$T_2 = \frac{L}{2} \mathbb{E}\|\eta_g \Delta^t - \eta_g v^t\|^2 = \frac{L\eta_g^2 \eta_s^2 R^2}{2} \mathbb{E}\|V^t\|^2. \tag{54}$$

From above T_2 expanding, with inequalities (22) and (23), we have inequality bound for $\mathbb{E}\|w^{t+1} - w^t\|^2$:

$$\mathbb{E}\|w^{t+1} - w^t\|^2 \leq \frac{4\eta_g^2 \eta_s^2}{K} \mathbb{E} \sum_{j \in [H_t]} \| \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t, r})\|^2 + \frac{\eta_g^2 \eta_s^2}{M} \mathbb{E} \sum_{j=1}^{M} \| \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\zeta_j^t, r})\|^2. \tag{55}$$

D.2.1. Combining $T_{1.2}$ and T_2

Thus, with $T_{1.2} + T_2$, we have

$$\begin{aligned}
&T_{1.2} + T_2 = \frac{L\eta_g^2 \eta_s^2 R^2 - \eta_g \eta_s R}{2} \mathbb{E}\|V^t\|^2 \\
&= \frac{L\eta_g^2 \eta_s^2 R^2 - \eta_g \eta_s R}{2} \mathbb{E}\| \frac{2}{KR} \sum_{j \in [H_t]} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t, r}) - \frac{1}{MR} \sum_{j=1}^{M} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\zeta_j^t, r})\|^2 \\
&\leq \frac{4(L\eta_g^2 \eta_s^2 R - \eta_g \eta_s)}{2K^2 R} \mathbb{E}\| \sum_{j \in [H_t]} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t, r})\|^2 + \frac{L\eta_g^2 \eta_s^2 R - \eta_g \eta_s}{2M^2 R} \mathbb{E}\| \sum_{j=1}^{M} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\zeta_j^t, r})\|^2.
\end{aligned} \tag{56}$$

where require η_s satisfying $\eta_s \leq \frac{\sqrt{2}}{\alpha N_j} \leq \frac{1}{8RL}$ from (38)

D.2.2. Expanding $T_{1.3}$

$$
\begin{aligned}
T_{1.3} = &\frac{\eta_g\eta_s R}{2}\mathbb{E}\|\frac{1}{M}\sum_{j=0}^{M}\nabla F(w^t) + \frac{1}{MR}\sum_{j=0}^{M}\sum_{r=0}^{R-1}\nabla F_j(w^t) - \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w^t) \\
&+ \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}(\nabla F_j(w^{t-\Gamma_j^t}) - \nabla F_j(w_j^{t-\Gamma_j^t,r})) - \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1}(\nabla F_j(w_j^{t-\zeta_j^t}) - \nabla F_j(w_j^{t-\zeta_j^t,r})) \\
&+ \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}(\nabla F_j(w^t) - \nabla F_j(w^{t-\Gamma_j^t})) - \frac{1}{MR}\sum_{j=1}^{M}\sum_{r=0}^{R-1}(\nabla F_j(w^t) - \nabla F_j(w_j^{t-\zeta_j^t}))\|^2 \\
\leq &\ \eta_g\eta_s R\mathbb{E}\|\frac{1}{M}\sum_{j=0}^{M}\nabla F(w^t) + \frac{1}{MR}\sum_{j=0}^{M}\sum_{r=0}^{R-1}\nabla F_j(w^t) - \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w^t)\|^2 \\
&+ \frac{2\eta_g\eta_s 4}{K}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla F_j(w^{t-\Gamma_j^t}) - \nabla F_j(w_j^{t-\Gamma_j^t,r})\|^2 \\
&+ \frac{2\eta_g\eta_s}{M}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla F_j(w_j^{t-\zeta_j^t}) - \nabla F_j(w_j^{t-\zeta_j^t,r})\|^2 \\
&+ \frac{2\eta_g\eta_s 4}{K}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla F_j(w^t) - \nabla F_j(w^{t-\Gamma_j^t})\|^2 \\
&+ \frac{2\eta_g\eta_s}{M}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\mathbb{E}\|(\nabla F_j(w^t) - \nabla F_j(w_j^{t-\zeta_j^t}))\|^2, \qquad (57)
\end{aligned}
$$

where last the inequalities holds by the $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Therefore, using smoothness (20) with the bound inequalities of (42) , (44) and (45), we can extend $T_{1.3}$,

$$
\begin{aligned}
T_{1.3} \leq &\ \eta_g\eta_s R\mathbb{E}\|\frac{1}{M}\sum_{j=0}^{M}\nabla F(w^t) + \frac{1}{MR}\sum_{j=0}^{M}\sum_{r=0}^{R-1}\nabla F_j(w^t) - \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w^t)\|^2 \\
&+ \frac{2\eta_g\eta_s 4L^2}{K}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\mathbb{E}\|w^{t-\Gamma_j^t} - w_j^{t-\Gamma_j^t,r}\|^2 + \frac{2\eta_g\eta_s L^2}{M}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\mathbb{E}\|w_j^{t-\zeta_j^t} - w_j^{t-\zeta_j^t,r}\|^2 \\
&+ \frac{2\eta_g\eta_s 4L^2}{K}\cdot RK\Gamma_{max}\sum_{q=t-\Gamma_j^t}^{t-1}\mathbb{E}\|w^{q+1} - w^q\|^2 + \frac{2\eta_g\eta_s L^2}{M}\cdot RM\zeta_{max}\sum_{q=t-\zeta_j^t}^{t-1}\mathbb{E}\|w^{q+1} - w^q\|^2, \qquad (58)
\end{aligned}
$$

where the inequality holds by applying the bound of parameter difference for the delayed sub-center model (44) and (45).

D.3. Expanding convergence inequality

By expanding and simplifying T_1 and T_2 in (47) with (38), (39), (23) and (22), using polarization identity $\langle a, b\rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2]$ with Eq. (29) of Δ_j^t, we have:

$$\begin{aligned}
&\mathbb{E}[F(w^{t+1})] - F(w^t) \\
&\leq -\frac{\eta_g \eta_s R}{2}\mathbb{E}\|\nabla F(w^t)\|^2 \\
&\quad + \eta_g \eta_s R \cdot \underbrace{\left(\quad \mathbb{E}\|\frac{1}{M}\sum_{j=0}^{M}\nabla F(w^t) + \frac{1}{MR}\sum_{j=0}^{M}\sum_{r=0}^{R-1}\nabla F_j(w^t) - \frac{2}{KR}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w^t)\|^2 \quad \right)}_{T_3} \\
&\quad + \frac{2\eta_g\eta_s 4L^2}{K}\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\mathbb{E}\|w^{t-\Gamma_j^t} - w_j^{t-\Gamma_j^t, r}\|^2 + \frac{2\eta_g\eta_s L^2}{M}\sum_{j=1}^{M}\sum_{r=0}^{R-1}\mathbb{E}\|w_j^{t-\zeta_j^t} - w_j^{t-\zeta_j^t, r}\|^2 \\
&\quad + 8\eta_g\eta_s R L^2 \Gamma_{max}\sum_{q=t-\Gamma_j^t}^{t-1}\mathbb{E}\|w^{q+1} - w^q\|^2 + 2\eta_g\eta_s R L^2\zeta_{max}\sum_{q=t-\zeta_j^t}^{t-1}\mathbb{E}\|w^{q+1} - w^q\|^2 \\
&+ \frac{4(L\eta_g^2\eta_s^2 R - \eta_g\eta_s)}{2K^2R}\mathbb{E}\|\sum_{j\in[H_t]}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\Gamma_j^t, r})\|^2 + \frac{L\eta_g^2\eta_s^2R - \eta_g\eta_s}{2M^2R}\mathbb{E}\|\sum_{j=1}^{M}\sum_{r=0}^{R-1}\nabla F_j(w_j^{t-\zeta_j^t, r})\|^2,
\end{aligned} \tag{59}$$

D.4. Expanding for T_3.

From the $T_{1.3}$ term above and bound diversity variance, with the same proof process of $T_{1.3}$ (58) we simplify T_3 and have

$$\begin{aligned}
T_3 &= \mathbb{E}\|\nabla F(w^t) + \nabla F(w^t) - \frac{1}{M}\sum_{j=0}^{M}(\nabla F(w^t) - \nabla F_j(w^t)) \\
&\quad - 2\nabla F(w^t) + \frac{2}{K}\sum_{j\in H_t}(\nabla F(w^t) - \nabla F_j(w^t))\|^2 \\
&\leq \frac{1}{M}\sum_{j=0}^{M}\mathbb{E}\|(\nabla F(w^t) - \nabla F_j(w^t))\|^2 + \frac{4}{K}\sum_{j\in[H_t]}\mathbb{E}\|(\nabla F(w^t) - \nabla F_j(w^t))\|^2 \\
&\leq \sigma_g^2 + \frac{4M}{K}\sigma_g^2 = \frac{K+4M}{K}\sigma_g^2,
\end{aligned} \tag{60}$$

where last inequality holds by variance diversity of partial models aggregation assumption of Lemma 1.

D.5. Simplify for convergence inequality.

With summing over $t = 1$ to T for both sides, we have

$$
\begin{aligned}
&\mathbb{E}[F(w^{T+1})] - F(w^1) \\
&\leq -\frac{\eta_g \eta_s R}{2} \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2 + \sum_{t=1}^{T} \eta_g \eta_s R \cdot \frac{K + 4M}{K} \sigma_g^2 + 8\eta_g \eta_s R L^2 (5R\eta_s^2(\sigma_s^2 + 6R \cdot \frac{M}{K}\sigma_g^2)T \\
&+ 30R^2\eta_s^2 \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2) + 2\eta_g \eta_s R L^2 (5R\eta_s^2(\sigma_s^2 + 6R\sigma_g^2)T + 30R^2\eta_s^2 \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2) \\
&+ 8\eta_g \eta_s R L^2 \Gamma_{max}^2 \sum_{t=1}^{T} \mathbb{E}\|w^{t+1} - w^t\|^2 + 2\eta_g \eta_s R L^2 \zeta_{max}^2 \sum_{t=1}^{T} \mathbb{E}\|w^{t+1} - w^t\|^2 \\
&+ \frac{4(L\eta_g^2\eta_s^2 R - \eta_g\eta_s)}{2K^2 R} \sum_{t=1}^{T} \mathbb{E}\| \sum_{j \in [H_t]} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\Gamma_j^t, r})\|^2 \\
&+ \frac{L\eta_g^2\eta_s^2 R - \eta_g\eta_s}{2M^2 R} \sum_{t=1}^{T} \mathbb{E}\| \sum_{j=1}^{M} \sum_{r=0}^{R-1} \nabla F_j(w_j^{t-\zeta_j^t, r})\|^2,
\end{aligned}
\tag{61}
$$

where the inequality holds by the bounds of accumulated gradients from partial sub-centers ($j \in H_t$) and full clients respectively at certain time step point through (38) and (39).

To simplify (59) with T_3, using (40) and Lemma G.1 in (Wang et al., 2023) with the bound of (55) and delay summation gradients bounds: $\sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^{t-\Gamma_j^t})\|^2 \leq \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2$ and $\sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^{t-\zeta_j^t})\|^2 \leq \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2$, then we reformulate (59) as:

$$
\begin{aligned}
&\mathbb{E}[F(w^{T+1})] - F(w^1) \\
&\leq -\frac{\eta_g \hat{\eta}_s R}{2} \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2 + \eta_g \hat{\eta}_s R \cdot \frac{K + 4M}{K} \sigma_g^2 T + 8\eta_g \hat{\eta}_s R L^2 (5R\hat{\eta}_s^2(\sigma_s^2 + 6R \cdot \frac{M}{K}\sigma_g^2)T \\
&+ 30R^2\hat{\eta}_s^2 \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2) + 2\eta_g \hat{\eta}_s R L^2 \cdot (5R\hat{\eta}_s^2(\sigma_s^2 + 6R\sigma_g^2)T + 30R^2\hat{\eta}_s^2 \sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2) \\
&+ (2\eta_g \hat{\eta}_s R L^2 \eta_g^2 \hat{\eta}_s^2 (4\Gamma_{max}^2 + \zeta_{max}^2) + \frac{L\eta_g^2\hat{\eta}_s^2 R - \eta_g\hat{\eta}_s}{2R}) \cdot \\
&\qquad (15R^3L^3\hat{\eta}_s^2 \cdot 4(\sigma_s^2 + 6R\frac{M}{K}\sigma_g^2)T + 15R^3L^3\hat{\eta}_s^2(\sigma_s^2 + 6R\sigma_g^2)T \\
&+ 3R^2(\frac{4M}{K} + 1)\sigma_g^2 T) \\
&+ (2\eta_g \hat{\eta}_s R L^2 \eta_g^2 \hat{\eta}_s^2 (4\Gamma_{max}^2 + \zeta_{max}^2) + \frac{L\eta_g^2\hat{\eta}_s^2 R - \eta_g\hat{\eta}_s}{2R}) \cdot (90R^4L^2\hat{\eta}_s^2 + 3R^2) \cdot 5 \sum_{t=1}^{T} \|\nabla F(w^t)\|^2.
\end{aligned}
\tag{62}
$$

In the inequality of above, we define $\hat{\eta}_s^2 = \frac{2+2L^2}{\alpha^2}$ and $\eta_s^2 \leq \hat{\eta}_s^2$, where we have established that the sub-center aggregation satisfies $\eta_s^2 \leq \frac{2+2L^2}{\alpha^2}$ from Inequality (36).

D.6. Analysis for Convergence Constraint and Leaning Rate

We extract coefficient from the terms which contain $\|\nabla F(w^t)\|^2$ from the above convergence inequality (62), and have:

$$C_{\nabla} = -\frac{\eta_g \hat{\eta}_s R}{2} + 10\eta_g \hat{\eta}_s R L^2 \cdot 30R^2 \hat{\eta}_s^2 + 5(2\eta_g \hat{\eta}_s R L^2 \eta_g^2 \hat{\eta}_s^2 (4\Gamma_{max}^2 + \zeta_{max}^2) \\ + \frac{L\eta_g^2 \hat{\eta}_s^2 R - \eta_g \hat{\eta}_s}{2R})(90R^4 L^2 \hat{\eta}_s^2 + 3R^2). \tag{63}$$

To ensure convergence and uphold the upper bound for inequality (62), we can have a constraint based on the third term on the right-hand side of (63):

$$2\eta_g \hat{\eta}_s R L^2 \eta_g^2 \hat{\eta}_s^2 (4\Gamma_{max}^2 + \zeta_{max}^2) + \frac{L\eta_g^2 \hat{\eta}_s^2 R - \eta_g \hat{\eta}_s}{2R} \leq 0. \tag{64}$$

From first and second terms on the right-hand side of (63), we also establish a constraint that satisfies :

$$10\eta_g \hat{\eta}_s R L^2 \cdot 30R^2 \hat{\eta}_s^2 \leq \frac{\eta_g \hat{\eta}_s R}{2}. \tag{65}$$

We then obtain:

$$\hat{\eta}_s^2 \leq \frac{1}{600R^2L^2}, \tag{66}$$

where the $\hat{\eta}_s$ (upper bound of η_s) holds the condition of constraint $\eta_s^2 \leq \frac{2+2L^2}{\alpha^2}$ and $\eta_s \leq \frac{1}{8RL}$ from (38). From bound of (66) and above constraints, we further obtain

$$\alpha \geq 10\sqrt{6(2+2L^2)}LR, \tag{67}$$

where α also holds the constraint $\alpha \geq 20L$ and $\alpha \geq 1$ from sub-center convergence rate (17).

Therefore we obtain the bound for η_g from above constraints

$$\eta_g \leq \frac{5\sqrt{6}(\sqrt{16(4\Gamma_{max}^2 + \zeta_{max}^2) + 1} - 1)}{4(4\Gamma_{max}^2 + \zeta_{max}^2)}. \tag{68}$$

D.7. General Convergence Rate

With the aforementioned constraints (67) and (68) where $\hat{\eta}_s^2 = \frac{2+2L^2}{\alpha^2}$, we exchange the terms on both sides of the inequality (62) and further simplify it. Then we consequently have general convergence rate:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla F(w^t)\|^2 \leq \frac{2\alpha}{C \cdot \eta_g RT}(F(w^1) - \mathbb{E}[F(w^{T+1})]) + 5RL^2(4-3L) \cdot \frac{BC^2}{\alpha^2 K}\sigma_s^2 \\ + 60R^3L^5\eta_g^2(4\Gamma_{max}^2 + \zeta_{max}^2) \cdot \frac{BC^4}{\alpha^4 K}\sigma_s^2 + 15R^2L^4\eta_g \cdot \frac{BC^3}{\alpha^3 K}\sigma_s^2 \\ + 12R^2L^2\eta_g^2(4\Gamma_{max}^2 + \zeta_{max}^2) \cdot \frac{(4M+K)C^2}{\alpha^2 K}\sigma_g^2 + 3RL\eta_g \cdot \frac{(4M+K)C}{\alpha K}\sigma_g^2, \tag{69}$$

where $C = \sqrt{2+2L^2}$ and $B = 5K + (24M + 6K)R$. This finishes the proof of Theorem 1. The $\alpha \geq 10\sqrt{6(2+2L^2)}LR$ from Ineq. (67) also hold the constrain $\alpha \geq 20L$ from the bound Ineq. (17).

D.8. Extending convergence analysis

Therefore, by choosing $\eta_g = \frac{1}{\sqrt{T}}$ and $\hat{\eta}_s = \sqrt{\frac{K}{R}}$, i.e., $\frac{2+2L^2}{\alpha^2} = \frac{K}{R}$ with $\alpha = \sqrt{\frac{(2+2L^2)R}{K}}$ from (69), the algorithm convergence rate has

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w^t)\|^2 \leq{}& \frac{2}{\sqrt{T}\sqrt{KR}}(F(w^1) - \mathbb{E}[F(w^{T+1})]) \\ &+ 5RL^2K(4-3L)\cdot\sigma_s^2\frac{5K+(24M+6K)R}{KR} \\ &+ 60R^3L^5\frac{1}{T}K^2(4\Gamma_{max}^2+\zeta_{max}^2)\sigma_s^2\frac{5K+(24M+6K)R}{KR^2} \\ &+ 15R^2L^4\frac{1}{\sqrt{T}}K\sqrt{K}\cdot\sigma_s^2\frac{5K+(24M+6K)R}{KR\sqrt{R}} \\ &+ 12R^2L^2\frac{1}{T}K(4\Gamma_{max}^2+\zeta_{max}^2)\cdot\sigma_g^2\frac{4M+K}{KR} \\ &+ 3RL\frac{1}{\sqrt{T}}\sqrt{K}\cdot\sigma_g^2\frac{4M+K}{K\sqrt{R}}. \qquad (70)\end{aligned}$$

Thus, we have the *proof* of Corollary 1. For nonconvex case, by choosing $\eta_g = \frac{1}{\sqrt{T}}$ and $\alpha = \sqrt{\frac{(2+2L^2)R}{K}}$, we can have convergence rate of HGA-FL algorithm 1 satisfies

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w^t)\|^2 ={}& \mathcal{O}\left(\frac{F(w^1)-F^*}{\sqrt{TKR}}\right) \\ &+ \mathcal{O}\left(\frac{60RL^5(4\Gamma_{max}^2+\zeta_{max}^2)KB\sigma_s^2}{T}\right) \\ &+ \mathcal{O}\left(\frac{12RL^2(4\Gamma_{max}^2+\zeta_{max}^2)(4M+K)\sigma_g^2}{T}\right) \\ &+ \mathcal{O}\left(\frac{15L^4\sqrt{KR}B\sigma_s^2}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{3\sqrt{R}L(4M+K)\sigma_g^2}{\sqrt{TK}}\right), \qquad (71)\end{aligned}$$

where $B = 5K + (24M + 6K)R$, and F^* is the optimal point of the objective. Thus, we have the *proof* of Corollary 1.

E. RELATED WORK

In recent years, FL methods such as SCAFFOLD (Karimireddy et al., 2020), FedProx (Li et al., 2020), and FedNova (Wang et al., 2020) have been proposed to mitigate the bias induced by data heterogeneity in distributed model aggregation. To enhance the efficiency of global model aggregation, recent works have introduced local regularization techniques and proximal objective function approximations (Li et al., 2020; Mishchenko et al., 2022; Malinovsky et al., 2022) to improve the consistency between local and global models. These approaches have demonstrated promising convergence properties for both nonconvex and convex objective functions.

Meanwhile, in the realm of asynchronous FL, several new methods (Yu et al., 2023; Wang et al., 2023) have been proposed to tackle gradient staleness and delays in heterogeneous device and data environments. A series of semi-asynchronous FL approaches (Sun et al., 2022) have also emerged, aiming to balance the bandwidth overhead of asynchronous aggregations on the server, the time overhead of synchronous updates, and the issue of model staleness.

Recent investigations into Multiple-Tier FL have primarily focused on three-tier architectures. Das et al. (Das and Patterson, 2021) and Malinovsky et al. (Malinovsky et al., 2022) configure each silo with a hub acting as a sub-aggregation node, subsequently transmitting the aggregated hub models to a global server for overall aggregation. FedHiSyn (Li et al., 2022) primarily groups devices into different tiers. Banerjee et al. (Banerjee et al., 2022) studied a three-tier model aggregation architecture and employed an L2 regularization term to maintain model personalization.

F. Explanations for the Assumptions

Assumption 1 is a common assumption adopted in convergence analysis for asynchronous FL (Wang et al., 2023). It indicates that $\zeta_j^t \geq \Gamma_j^t$ at any global step t which also implies $\zeta_{max} \geq \Gamma_{max}$. According to Assumption 1, $\Delta_j^{t-\Gamma_j^t}$ represents the parameters difference in the model updates for sub-center j from the step point $t - \Gamma_j^t$ when sub-center j starts to compute its internal gradients with its clients.

In FL DNNs training, it often commits the nonconvex objective in real-time model learning process. This convergence process is affected by iterations of update, smoothness constant L, local gradient variance bound, and learning rate. We present our results with following common assumptions in FL (Wang et al., 2020; Reddi et al., 2021; Toghani and Uribe, 2022). Assumptions 3 also implies $\|\nabla F_j(w) - \nabla F_j(w')\| \leq L\|w - w'\|$ from Equation 1 due to averaging local objective.

F.1. Proof for Assumptions 6

For Assumption 6 (Bounded variance across sub-centers). The variance of stochastic gradients in each sub-center $j \in [m]$ is bounded, and satisfies:

$$\mathbb{E}_{\xi_j \sim D_j}[\|\nabla f_j(w;\xi_j) - \nabla F_j(w)\|^2] \leq \hat{\sigma}_j^2, \tag{72}$$

where $\hat{\sigma}_j^2 = \max_{i\in[n_j]} \hat{\sigma}_{j,i}^2$ and $\nabla f_j(w;\xi_j)$ denotes the average stochastic gradient on sub-center j for data point set D_j which belongs to j. $[m]$ represents the sub-center index set.

Proof

$$\begin{aligned}
&\mathbb{E}_{\xi_j\sim D_j}[\|\nabla f_j(w;\xi_j) - \nabla F_j(w)\|^2] \\
&= \mathbb{E}_{\xi_j\sim D_j}[\|\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla f_{j,i}(w;\xi_{j,i}) - \frac{1}{N_j}\sum_{i=1}^{N_j}\nabla F_{j,i}(w)\|^2] \\
&\le \frac{1}{N_j}\sum_{i=1}^{N_j}[\mathbb{E}_{\xi\sim D_{j,i}}\|\nabla f_{j,i}(w;\xi_{j,i}) - \nabla F_{j,i}(w)\|^2] \\
&\le \max_{i\in[n_j]} \hat{\sigma}_{j,i}^2 = \hat{\sigma}_j^2.
\end{aligned}$$

■

G. Detail of Experiment Setting

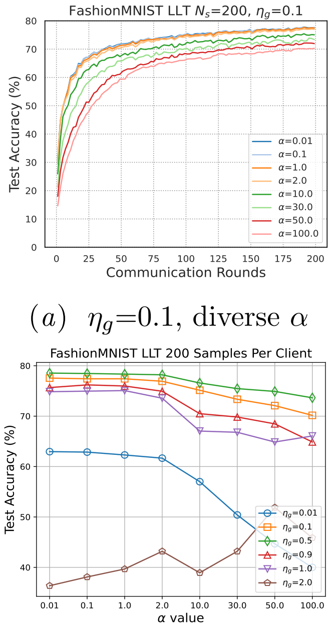
Datasets and Models. We evaluated on EMNIST, FashionMNIST (F-MNIST) and CIFAR-10 datasets. EMNIST has 47 classes, 112,800 training and 18,800 test samples. For non-i.i.d. and imbalanced data, we use Label Dirichlet Allocation (LDA) and local long-tailed (LLT) partitioning (Tang et al., 2021). LDA draws samples from $\mathrm{Dir}(\alpha_d)$ per client, where α_d is the Dirichlet concentration factor. We adopt LLT $\alpha_l = 0.9$, with one class occupying 90% of samples. For LLT in experiments, we default to using 200 samples per client ($N_s = 200$). We train with the classical Two-Convolution Layers (2-Conv) DNN suggested in many works (Das and Patterson, 2021) and Resnet-18 (Jhunjhunwala et al., 2023; Acar et al., 2021). Except for CIFAR-10, without a specific statement, we use 2-Conv for all datasets.

Staleness Setting. Practically, we use global time step $\hat{t}$ to serve as the foundational time unit for the entire system. This time step derives from the average computational time required for mini-batch training across all local workers. Throughout training, each local worker executes a specific number of global time steps $\hat{t}$ based on its capabilities. The duration between consecutive global fusion steps t and $t+1$ depends on $\hat{t}$, relative to the average time taken for mini-batch gradient computation across all workers. Without specific statement, we set $\Gamma_{\max} = 500$ as default, corresponding to 500 units of $\hat{t}$.

Diverse Multi-center FLs. We propose a set of two-level hierarchical joint aggregation FL methods combining asynchronous and synchronous approaches. At the upper level, we employ asynchronous global aggregation using FedAsync, FedBuff and CA^{2}FL baselines, denoted as FedAsync-G, FedBuff-G, and CA^{2}FL-G, respectively. At the lower level, we employ synchronous sub-center aggregation with FedAvg (S-Avg), FedProx (S-Prox), and FedDyn (S-Dyn) baselines. In these original single-center methods, clients perform normal gradient updates, which we feasibly substitute with sub-center model update processes.

We employ the global aggregation component of HGA-FL from Algorithm 1 Line 3 to Line 16 denoted as HGA-FL-G, alongside diverse sub-center fusion methods to assess their performance. S-Avg, an example of sub-center synchronous fusion within the hierarchy framework, is depicted in Algorithm 3 from the Supplementary Material.

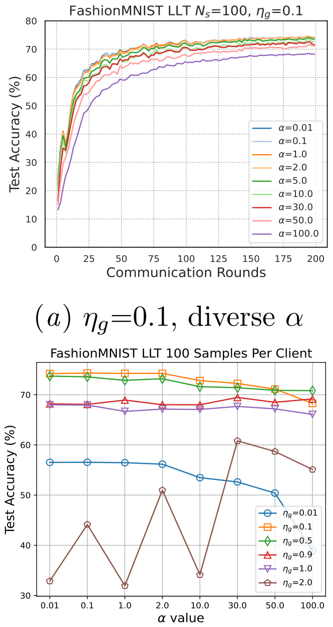
Common Settings. For compared methods, we adopt default hyperparameters from their original works for both asynchronous and synchronous approaches. S-Dyn uses the same value of α as HGA-FL's sub-center aggregation. For S-Prox, we set $\mu = 2$ (Li et al., 2020) consistent with the same coefficient in HGA-FL's regularization term. For FedBuff-G, we use a decay rate for the step size η_g in (Nguyen et al., 2022) for stability. For HGA-FL, unless stated otherwise, we default to $\eta_g = 0.1$, $\alpha = 2$. All other common settings are $M = 8$, $K = 3$, $\Gamma_{\max} = 500$, $R = 8$, $T = 200$, LDA $\alpha = 0.2$ with 10% sample quantity, 50 clients, 2 local epochs for global test accuracy.



(a) η_g=0.1, diverse α

(b) Sensitivity of α and η_g

Figure 6: 100 clients, 8 sub-centers and $N_s = 200$.



(a) η_g=0.1, diverse α

(b) Sensitivity of α and η_g

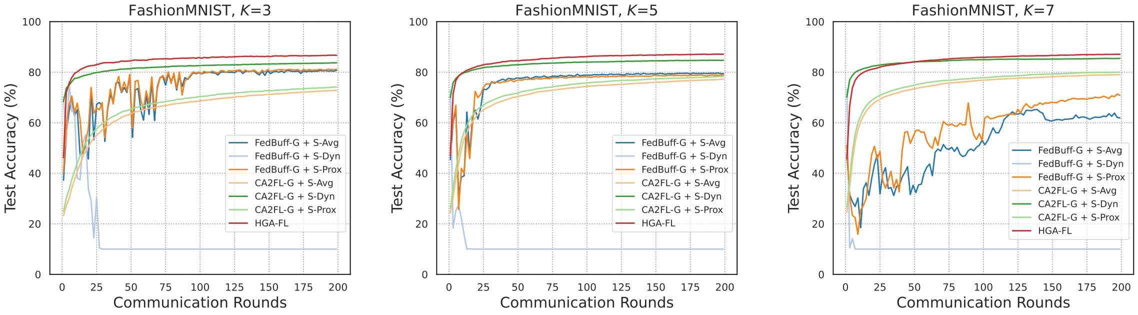
Figure 7: 32 clients, 8 sub-centers and $N_s = 100$.

H. Detail Result and Analysis

In this section, we experimentally compare the efficiency and extendibility of our hierarchical HGA-FL with other combined asynchronous and synchronous FL methods.

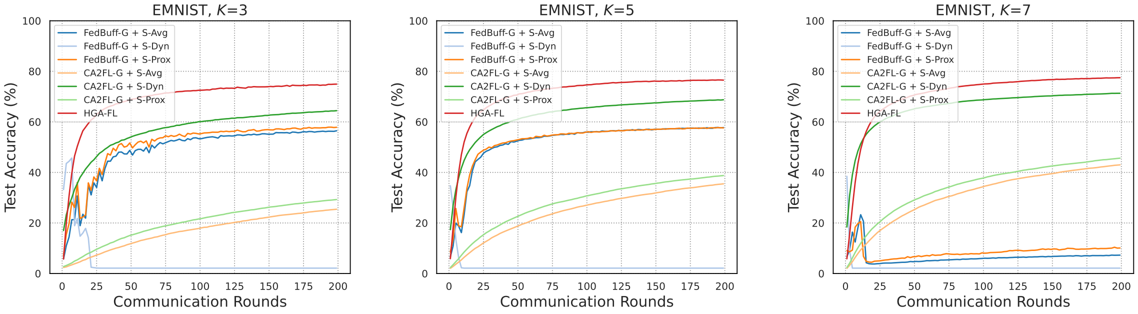
Effect of η_g and α. To evaluate hyperparameter sensitivity, we utilize the FashionMNIST dataset in HGA-FL, varying η_g and α for global and sub-center aggregation, respectively. We employ 100 clients with 5 local epochs, $R = 5$, $N_s = 200$, LLT α_l=0.9, a 2-Conv DNN, $K = 4$, $M = 8$ and $\Gamma_{max} = 500$ relative to $\hat{t}$ for experimentation. We also conduct another experiment using 32 clients and $N_s = 100$ to observe different behavior. The results are depicted in Figure 6 and Figure 7. Our findings indicate that, η_g value ranging from 0.1

to 0.5 and the α values ranging from 1 to 2 demonstrate optimal performance, consistent with the constraints outlined in Theorem 1 and Corollary 1 for a certain L value.



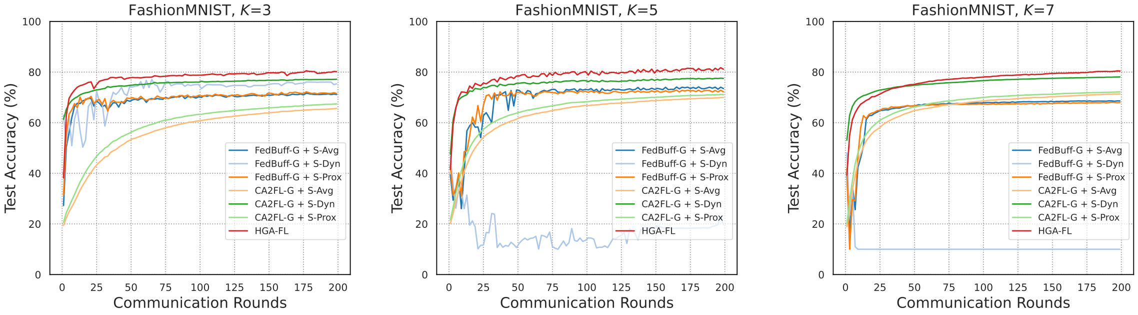
(a) $K = 3$ and 100 clients (b) $K = 5$ and 100 clients (c) $K = 7$ and 100 clients

Figure 8: Diverse MC-FLs via LDA $\alpha_d = 0.2$, 100% F-MNIST samples, $M = 8$ and $R = 8$.



(a) $K = 3$ and 100 clients (b) $K = 5$ and 100 clients (c) $K = 7$ and 100 clients

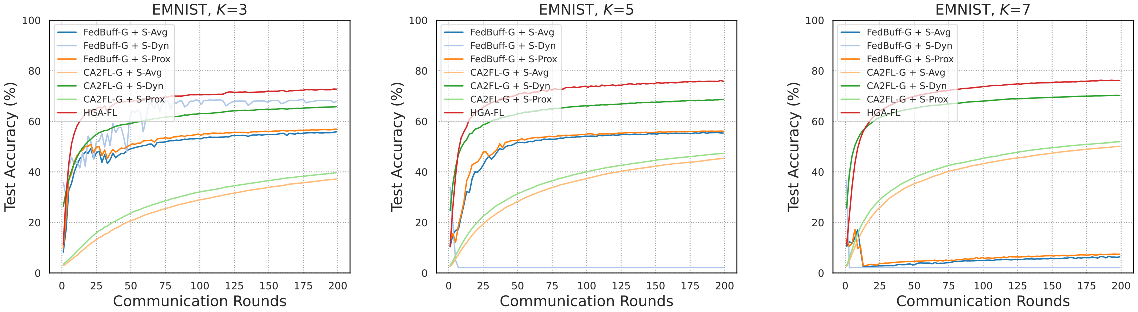
Figure 9: Diverse MC-FLs via LDA $\alpha_d = 0.2$, 10% EMNIST samples, $M = 8$ and $R = 8$.



(a) $K = 3$ and 50 clients (b) $K = 5$ and 50 clients (c) $K = 7$ and 50 clients

Figure 10: Diverse MC-FLs via LDA $\alpha_d = 0.2$, 10% F-MNIST samples, $M = 8$ and $R = 8$.

Effects of Diverse Global Asynchronous methods with Buffer. For comparing with baseline FL multiple tiers conjunction methods, we construct a set of 3-tier architecture



(a) $K = 3$ and 50 clients (b) $K = 5$ and 50 clients (c) $K = 7$ and 50 clients

Figure 11: Diverse MC-FLs via LDA $\alpha_d = 0.2$, 10% EMNIST samples, $M = 8$ and $R = 8$.

of two level aggregation combination methods. On the global asynchronous aggregation level, we adopt asynchronous baseline global aggregation methods FedBuff-G, and CA^{2}FL-G. Next, we incorporate sub-center synchronous baseline S-Avg, S-Prox and S-Dyn. We compare these combined multi-center FL (MC-FL) methods with our HGA-FL methods based on global model accuracy, which indicates generalization. We adopt $\eta_g = 0.1$ and $\alpha = 2$ with LDA $\alpha_d = 0.2$ and common settings, utilizing 100% of the dataset samples for FashionMNIST and 10% of the dataset samples for EMNIST. The results are shown in Table 2. Detailed figures are presented in Figure 8 and Figure 9. We also conducted additional experiments with 50 clients, each using 10% of the dataset samples. The results of these experiments can be found in Figure 10 for FashionMNIST and Figure 11 for EMNIST. We observe that a large buffer size K helps improve global test accuracy and that our method outperforms other combined algorithms in these settings, indicating its strong generalization capabilities.

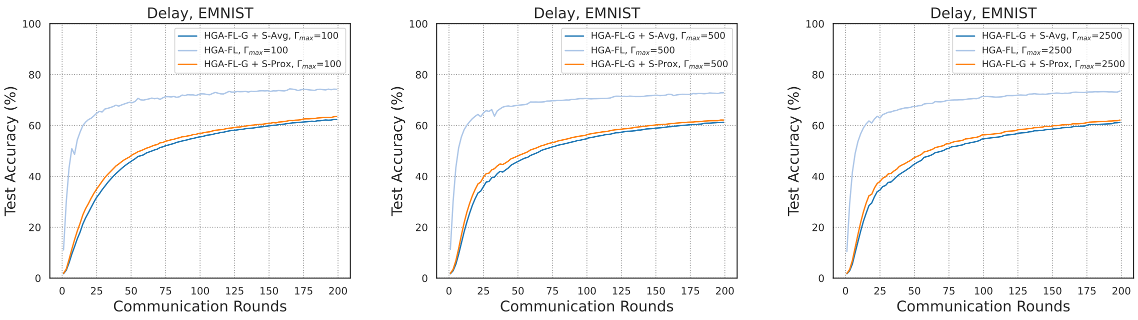
The baseline FedAsync adopt a proximity regularization term in the client update which similar to FedProx. In the experiment, We adopt the FedAsync global fusion method denoted by FedAsync-G, and S-FedProx in the sub-center.

Effects of HGA-FL-G with Diverse Sub-center Aggregation. We compare HGA-FL to an integrated version HGA-FL-G incorporating S-Avg and S-Prox sub-center aggregations, using LDA $\alpha_d = 0.2$. We also evaluate SCAFFOLD-ExP (Jhunjhunwala et al., 2023) for sub-center aggregation (S-SCAFF-ExP), an optimized SCAFFOLD version with default settings in their original paper. However, SCAFFOLD has been reported unstable under staleness and data imbalance in many previous works (Reddi et al., 2021; Yu et al., 2022). We observed similar phenomena with S-SCAFF-ExP in hierarchical FL. The results are shown in Tables 3 and 4. And the results in Table 4 show HGA-FL outperforms others on EMNIST and F-MNIST datasets.

Compare to Asynchronous and Synchronous Global Aggregation. We compare our method with both vanilla asynchronous and synchronous global aggregation methods, each combined with different sub-center fusion methods. For global methods, we adopt FedAsync global fusion (FedAsync-G) (Xie et al., 2019) and synchronous averaging (Sync Avg) methods. Results are presented in Table 3 and Figure 5. Notably, multi-center FLs with global synchronous method, the Sync Avg, require more time steps $\hat{t}$ to achieve the same 54% accuracy compared to asynchronous global methods. While HGA-FL outperforms other methods overall, FedAsync-G with S-Dyn exhibits faster convergence. Despite lower speed metrics, its overall accuracy within 200 global rounds surpasses FedAsync-G. However, FedAsync-G's training curve displays significant fluctuations and instability (see Figure 5 (a)).

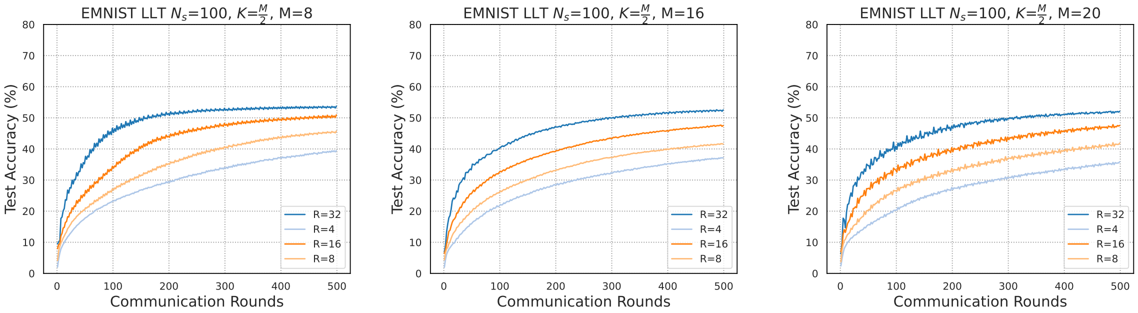
Compare Across Models. We compared HGA-FL to global methods (FedBuff-G and CA^{2}FL-G) with S-Dyn on both 2-Conv DNN and ResNet18 models under two LDA distributions and with a 10% sample quantity. S-Dyn exhibited superior performance compared to other sub-center methods in previous experiments, so we exclusively used S-Dyn in the sub-center for this experiment. Results from Table 5 indicate that HGA-FL maintains its superiority over most alternatives. Particularly, multi-center FLs training the ResNet18 model with deeper layers exhibit higher test accuracy. These findings highlight the strong generalization capabilities of the HGA-FL method among asynchronous global methods.

Effect of Staleness. We conducted a comparison between HGA-FL and HGA-FL-G using S-Avg and S-Prox with different maximum delays Γ_{max} under EMNIST and LDA $\alpha = 0.2$. Results from Table 6 and Figure 12 in the Supplementary Material show that a larger Γ_{max} indeed affects the model's test accuracy. However, HGA-FL continues to outperform, even with $\Gamma_{max} = 2500$.



(a) $\Gamma_{max} = 100$ (b) $\Gamma_{max} = 500$ (c) $\Gamma_{max} = 2500$

Figure 12: Effect of staleness.



(a) $M = 8$ (b) $M = 16$ (c) $M = 20$

Figure 13: Effect of M and R

Effect of M and R. We explore various combinations of sub-center numbers M and aggregation rounds R in HGA-FL comparisons. Using a buffer size K equal to half the value of M, we deploy 100 clients on the EMNIST dataset following LLT distribution $N_s = 100$ and $\alpha_l = 0.9$. These experiments are conducted over $T = 500$ rounds with common settings. Results in Table 7 reveal that a smaller number of sub-centers M with corresponding buffer size K achieve a higher global model test accuracy. This suggests that a larger number of sub-centers engaged in fusion with the same total number of clients may introduce more gradient variance and model drift, ultimately reducing global test accuracy. These findings are consistent with the convergence rates specified in Theorem 1 and Corollary 1 containing terms related to M and K. More results figures can be found in Figure 13.

I. Extension of Algorithm

The algorithm 3 is the vanilla model averaging method for sub-center internal aggregation.

Algorithm 3 Sub-centers Averaging Aggregation (S-Avg)

```
Input: $w^1$, $\eta_l$, $\{n_i\}_{i=1}^{N_j}$, $n$,$[m]$,$\{n_j\}_{j=1}^{j=|[m]|}$;
S-Avg Sub-centers Procedure:
for each $j \in [m]$ Sub-center $j$ in parallel do
    $t_j \leftarrow 1$, $w_j^0 \leftarrow w^1$
    repeat
        if Global $w^t$ update then
            Receive $w^t$; $w_j^{t_j-1} \leftarrow w^t$ asynchronously
            Clients $\forall i \in [n_j]$ , $w_{j,i}^{t_j} \leftarrow w_j^{t_j-1}$
        end if
        for client $i \in [n_j]$ in parallel do
            for local step k from 1 to E do
                $w_{j,i}^{t_j} = w_{j,i}^{t_j} - \eta_l \nabla F_{j,i}(w_{j,i}^{t_j})$
            end for
            Transmit client $w_{j,i}^{t_j}$ to sub-center $j$
        end for
        $w_j^{t_j} = \sum_{i=1}^{|n_j|} \frac{n_i}{n} w_{j,i}^{t_j}$
        if $t_j == R$ then
            Evaluate $\Gamma_j^t$ , $\Delta_j^{t-\Gamma_j^t} = w_j^{t_j} - w_j^{t_j-R}$
            Transmit $\Delta_j^{t-\Gamma_j^t}$ to Global Server, $t_j \leftarrow 1$
        end if
        $t_j \leftarrow t_j + 1$
    until Global Server stop
end for
```