Appendix A: Convergence Proof for Full Client Participation Under Convex Condition

A.1. Proof of Lemmas

Proof of Lemma 1. In section 4.2, we have $\bar{\theta}_{t+1} = \bar{\theta}_t - \eta_t g_t$, then

$$\begin{aligned}\left\|\overline{\theta}_{t+1} - \theta^*\right\|^2 &= \left\|\overline{\theta}_t - \eta_t g_t - \theta^*\right\|^2 = \left\|\overline{\theta}_t - \theta^* - \eta_t \overline{g}_t + \eta_t \overline{g}_t - \eta_t g_t\right\|^2 \\ &= \underbrace{||\overline{\theta}_t - \theta^* - \eta_t \overline{g}_t||^2}_{A_1} + \underbrace{\eta_t^2 ||\overline{g}_t - g_t||^2}_{A_2} + \underbrace{2\eta_t \langle \overline{\theta}_t - \theta^* - \eta_t \overline{g}_t, \overline{g}_t - g_t \rangle}_{A_3}.\end{aligned} \tag{A.1}$$

First, we bound the term A_1 as follows: $A_1 = ||\overline{\theta}_t - \theta^* - \eta_t \overline{g}_t||^2 = ||\overline{\theta}_t - \theta^*||^2 + \underbrace{\eta_t^2 ||\overline{g}_t||^2}_{B_1} - \underbrace{2\eta_t \langle \overline{\theta}_t - \theta^*, \overline{g}_t \rangle}_{B_2}$. To further bound the term B_1, we have:

$$B_1 = \eta_t^2 ||\overline{g}_t||^2 = \eta_t^2 \left\| \sum_{k=1}^K \varphi_{t,k} \nabla \ell_k(\theta_{t,k}) \right\|^2 \overset{(a1)}{\leq} \eta_t^2 \sum_{k=1}^K \varphi_{t,k} \left\| \nabla \ell_k(\theta_{t,k}) \right\|^2 \overset{(a2)}{\leq} 2L\eta_t^2 \sum_{k=1}^K \varphi_{t,k} \left[\ell_k(\theta_{t,k}) - \ell_k^* \right]. \tag{A.2}$$

where $(a1)$ is the Cauchy–Schwarz inequality $\left\|\sum_{i=1}^N a_i * b_i\right\|^2 \leq \sum_{i=1}^N {a_i}^2 * \sum_{i=1}^N {b_i}^2$. $(a2)$ is due to the L-smoothness property of $\ell_k(\cdot)$: $\left\|\nabla \ell_k(\theta_{t,k}) - \nabla \ell_k^*\right\|^2 \leq 2L \left[\ell_k(\theta_{t,k}) - \ell_k^* \right]$. To further bound the term B_2, we have:

$$\begin{aligned}B_2 &= 2\eta_t \langle \overline{\theta}_t - \theta^*, \overline{g}_t \rangle = 2\eta_t \left\langle \overline{\theta}_t - \theta^*, \sum_{k=1}^K \varphi_{t,k} \nabla \ell_k(\theta_{t,k}) \right\rangle = 2\eta_t \sum_{k=1}^K \varphi_{t,k} \left\langle \overline{\theta}_t - \theta_{t,k} + \theta_{t,k} - \theta^*, \nabla \ell_k(\theta_{t,k}) \right\rangle \\ &= \underbrace{2\eta_t \sum_{k=1}^K \varphi_{t,k} \left\langle \overline{\theta}_t - \theta_{t,k}, \nabla \ell_k(\theta_{t,k}) \right\rangle}_{C_1} + \underbrace{2\eta_t \sum_{k=1}^K \varphi_{t,k} \left\langle \theta_{t,k} - \theta^*, \nabla \ell_k(\theta_{t,k}) \right\rangle}_{C_2}.\end{aligned} \tag{A.3}$$

To further bound the term C_1, we get:

$$\begin{aligned}-C_1 &= -2\eta_t \sum_{k=1}^K \varphi_{t,k} \left\langle \overline{\theta}_t - \theta_{t,k}, \nabla \ell_k(\theta_{t,k}) \right\rangle \overset{(a3)}{\leq} \sum_{k=1}^K \varphi_{t,k} \left\| \overline{\theta}_t - \theta_{t,k} \right\|^2 + \eta_t^2 \sum_{k=1}^K \varphi_{t,k} \left\| \nabla \ell_k(\theta_{t,k}) \right\|^2 \\ &\overset{(a4)}{\leq} \sum_{k=1}^K \varphi_{t,k} \left\| \overline{\theta}_t - \theta_{t,k} \right\|^2 + 2L\eta_t^2 \sum_{k=1}^K \varphi_{t,k} \left[\ell_k(\theta_{t,k}) - \ell_k^* \right],\end{aligned} \tag{A.4}$$

where $(a3)$ is an inequality: $2\langle a, b \rangle \leq \lambda ||a||^2 + \frac{1}{\lambda} ||b||^2, \lambda > 0$, so $-2 \left\langle \overline{\theta}_t - \theta_{t,k}, \nabla \ell_k(\theta_{t,k}) \right\rangle \leq \lambda \left\| \overline{\theta}_t - \theta_{t,k} \right\|^2 + \frac{1}{\lambda} \left\| \nabla \ell_k(\theta_{t,k}) \right\|^2 and\ \lambda = \frac{1}{\eta_t} > 0$, $(a4)$ is the property of L-smoothness: $\left\| \nabla \ell_k(\theta_{t,k}) - \nabla \ell_k^* \right\|^2 \leq 2L \left[\ell_k(\theta_{t,k}) - \ell_k^* \right]$.

Then, in order to further bound the term C_2, we obtain:

$$-C_2 = -2\eta_t \sum_{k=1}^K \varphi_{t,k} \left\langle \theta_{t,k} - \theta^*, \nabla \ell_k(\theta_{t,k}) \right\rangle \overset{(a5)}{\leq} -2\eta_t \sum_{k=1}^K \left\{ \varphi_{t,k} \left[\ell_k(\theta_{t,k}) - \ell_k(\theta^*) \right] \right\} - \mu\eta_t \sum_{k=1}^K \varphi_{t,k} \left\| \theta_{t,k} - \theta^* \right\|^2, \tag{A.5}$$

where (a_5) is the $\mu - strong$ convexity of $\ell_k(\theta_{t,k}) : -\left\langle x - y, \nabla \ell_k(x) \right\rangle \leq -\left(\ell_k(x) - \ell_k(y) \right) - \frac{\mu}{2} \left\| x - y \right\|^2$.

With the above results of the term C_1 and C_2, we have:

$$\begin{aligned}B_2 = C_1 + C_2 \geq &- \left\{ \sum_{k=1}^K \varphi_{t,k} \left\| \overline{\theta}_t - \theta_{t,k} \right\|^2 + 2L\eta_t^2 \sum_{k=1}^K \varphi_{t,k} \left[\ell_k(\theta_{t,k}) - \ell_k^* \right] \right\} \\ &- \left\{ -2\eta_t \sum_{k=1}^K \left\{ \varphi_{t,k} [\ell_k(\theta_{t,k}) - \ell_k(\theta^*)] \right\} - \mu\eta_t \sum_{k=1}^K \varphi_{t,k} \left\| \theta_{t,k} - \theta^* \right\|^2 \right\}.\end{aligned} \tag{A.6}$$

By combining B_1, B_2, we have:

$$\begin{aligned}
A_1 =& \leq ||\overline{\theta}_t - \theta^*||^2 + 2L\eta_t^2 \sum_{k=1}^{K} \varphi_{t,k} \left[\ell_k(\theta_{t,k}) - \ell_k^*\right] + \left\{ \sum_{k=1}^{K} \varphi_{t,k} \left\|\overline{\theta}_t - \theta_{t,k}\right\|^2 + 2L\eta_t^2 \sum_{k=1}^{K} \varphi_{t,k} [\ell_k(\theta_{t,k}) - \ell_k^*] \right\} \\
&+ \left\{ -2\eta_t \sum_{k=1}^{K} \{\varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\theta^*)]\} - \mu\eta_t \sum_{k=1}^{K} \varphi_{t,k} \left\|\theta_{t,k} - \theta^*\right\|^2 \right\} \\
&= (1 - \mu\eta_t)||\overline{\theta}_t - \theta^*||^2 + \underbrace{4L\eta_t^2 \sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k^*] - 2\eta_t \sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\theta^*)]}_{D} + \sum_{k=1}^{K} \varphi_{t,k} \left\|\overline{\theta}_t - \theta_{t,k}\right\|^2.
\end{aligned}$$

Now, to further bound the term D, we have: $D = 2\eta_t(2L\eta_t - 1)\underbrace{\sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\theta^*)]}_{E} + 4L\eta_t^2[\ell(\theta^*) - \sum_{k=1}^{K} \varphi_{t,k}\ell_k^*]$. Next, let's define $\mathcal{Q} = \ell(\theta^*) - \sum_{k=1}^{K} \varphi_{t,k}\ell_k^*$. Consequently, we can express D as follows:

$$D = 2\eta_t(2L\eta_t - 1)\underbrace{\sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\theta^*)]}_{E} + 4L\eta_t^2\mathcal{Q}. \tag{A.7}$$

Then, to further bound the term E.

$$\begin{aligned}
E =& \sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\overline{\theta}_t) + \ell_k(\overline{\theta}_t) - \ell_k(\theta^*)] = \sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\theta_{t,k}) - \ell_k(\overline{\theta}_t)] + \sum_{k=1}^{K} \varphi_{t,k}[\ell_k(\overline{\theta}_t) - \ell_k(\theta^*)] \\
&\overset{(a6)}{\geq} \sum_{k=1}^{K} \varphi_{t,k} \left\langle \nabla\ell_k(\overline{\theta}_t), \theta_{t,k} - \overline{\theta}_t \right\rangle + [\ell(\overline{\theta}_t) - \ell(\theta^*)] \\
&\overset{(a7)}{\geq} -\frac{\eta_t}{2} \sum_{k=1}^{K} \varphi_{t,k} \left\|\nabla\ell_k(\overline{\theta}_t)\right\|^2 - \frac{1}{2\eta_t} \sum_{k=1}^{K} \varphi_{t,k} \left\|\theta_{t,k} - \overline{\theta}_t\right\|^2 + [\ell(\overline{\theta}_t) - \ell(\theta^*)] \\
&\overset{(a8)}{\geq} -L\eta_t \sum_{k=1}^{K} \varphi_{t,k}(\ell_k(\overline{\theta}_t) - \ell_k^*) - \frac{1}{2\eta_t} \sum_{k=1}^{K} \varphi_{t,k} \left\|\theta_{t,k} - \overline{\theta}_t\right\|^2 + [\ell(\overline{\theta}_t) - \ell(\theta^*)].
\end{aligned} \tag{A.8}$$

Where $(a6)$ is the convexity of ℓ_k, $(a7)$ is an inequality $2\langle a, b\rangle \leq \lambda||a||^2 + \frac{1}{\lambda}||b||^2, \lambda > 0$, resulting in: $-2\langle \overline{\theta}_t - \theta_{t,k}, \nabla\ell(\theta_{t,k})\rangle \leq \lambda||\overline{\theta}_t - \theta_{t,k}||^2 + \frac{1}{\lambda}||\nabla\ell(\theta_{t,k})||^2, \lambda = \frac{1}{\eta_t} > 0$, $(a8)$ signifies the use of the L-smoothness property. Then,

$$\begin{aligned}
D &\leq (1 - 2L\eta_t) \sum_{k=1}^{K} \varphi_{t,k} \left\|\theta_{t,k} - \overline{\theta}_t\right\|^2 + 2\eta_t(2L\eta_t - 1)(1 - L\eta_t)[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q} \\
&\leq (1 - 2L\eta_t) \sum_{k=1}^{K} \varphi_{t,k} \left\|\theta_{t,k} - \overline{\theta}_t\right\|^2 - \frac{3}{4}\eta_t[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q}.
\end{aligned} \tag{A.9}$$

We have already defined $\mathcal{Q} = \ell(\theta^*) - \sum_{k=1}^{K} \varphi_{t,k}\ell_k^*$. Given that $\eta_t \leq \frac{1}{4L}$, we have $L\eta_t \leq \frac{1}{4}$ and $(2L\eta_t - 1)(1 - L\eta_t) \leq -\frac{3}{8}$. Now, we can obtain A_1 as follows:

$$\begin{aligned}
A_1 &= (1 - \mu\eta_t) \left\|\overline{\theta}_t - \theta^*\right\|^2 + (1 - 2L\eta_t) \sum_{k=1}^{K} \varphi_{t,k}||\theta_{t,k} - \overline{\theta}_t||^2 - \frac{3}{4}\eta_t[\ell(\overline{\theta}_t) - \ell(\theta^*)] \\
&\quad + 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q} + \sum_{k=1}^{K} \varphi_{t,k} \left\|\overline{\theta}_t - \theta_{t,k}\right\|^2 \\
&= (1 - \mu\eta_t) \left\|\overline{\theta}_t - \theta^*\right\|^2 + 2(1 - L\eta_t) \sum_{k=1}^{K} \varphi_{t,k}||\overline{\theta}_t - \theta_{t,k}||^2 - \frac{3}{4}\eta_t[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q}.
\end{aligned} \tag{A.10}$$

Hence,

$$\left\|\overline{\theta}_{t+1}-\theta^*\right\|^2 = A_1+A_2+A_3 = (1-\mu\eta_t)\left\|\overline{\theta}_t-\theta^*\right\|^2 + 2(1-L\eta_t)\sum_{k=1}^{K}\varphi_{t,k}\left\|\overline{\theta}_t-\theta_{t,k}\right\|^2 - \frac{3}{4}\eta_t[\ell(\overline{\theta}_t)-\ell(\theta^*)]$$
$$+ 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q} + \eta_t^2||\overline{g}_t-g_t||^2 + 2\eta_t\left\langle\overline{\theta}_t-\theta^*-\eta_t\overline{g}_t, \overline{g}_t-g_t\right\rangle.$$

Proof of Lemmas 2 Notice that g_t and $\overline{g}_t$are defined in 4.2 and we have:

$$\begin{aligned}\mathbb{E}\left\|g_t-\overline{g}_t\right\|^2 &= \mathbb{E}\left\|\sum_{k=1}^{K}\varphi_{t,k}\nabla\ell_k(\theta_{t,k},\xi_t^{k,i}) - \sum_{k=1}^{K}\varphi_{t,k}\nabla\ell_k(\theta_{t,k})\right\|^2 \\ &= \sum_{k=1}^{K}\varphi_{t,k}^2\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i}) - \nabla\ell_k(\theta_{t,k})\right\|^2 = \sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2.\end{aligned} \tag{A.11}$$

The final step is due to the Assumption 3, which bounds the variance of the stochastic gradients for client k by σ_k^2.

Proof of Lemmas 3. Each time client k receives the model parameters from the server, it performs local updates for E steps before sending the locally updated model back to the server. Let t_0 denote the starting time at which the server sends the model parameters to client k for updating.

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\varphi_{t,k}\left\|\overline{\theta}_t-\theta_{t,k}\right\|^2\right] = \sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|\overline{\theta}_t-\theta_{t,k}\right\|^2 = \sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|(\theta_{t,k}-\overline{\theta}_{t0})-(\overline{\theta}_t-\overline{\theta}_{t0})\right\|^2 \\ &\overset{(b1)}{\leq}\sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|\theta_{t,k}-\overline{\theta}_{t0}\right\|^2 - \left\|\mathbb{E}(\overline{\theta}_t-\overline{\theta}_{t0})\right\|^2 \leq \sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|\theta_{t,k}-\overline{\theta}_{t0}\right\|^2 \\ &\leq \sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\sum_{t=t_0}^{t_0+E-1}\eta_t^2\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2 \overset{(b2)}{\leq}\sum_{k=1}^{K}\varphi_{t,k}\eta_{t_0}^2E^2G^2 \overset{(b3)}{\leq}\sum_{k=1}^{K}\varphi_{t,k}4\eta_t^2E^2G^2 = 4\eta_t^2E^2G^2,\end{aligned} \tag{A.12}$$

where $(b1)$ uses the expression: $\mathbb{E}\left\|X-EX\right\|^2 = \mathbb{E}\left\|X\right\|^2 - \left\|\mathbb{E}X\right\|^2$, X represents $(\theta_{t,k}-\overline{\theta}_{t0})$ in (A.12). $(b2)$ includes the inequality $||\sum_{t=1}^{E}a_t||^2 \leq E\sum_{t=1}^{E}||a_t||^2$, $\eta_t \leq \eta_{t_0}$ for $t \geq t_0$ (which implies that η_t is non-increasing), and the assumption 4: $\mathbb{E}\left\|\nabla\ell_k(\theta_t^k,\xi_t^{k,i})\right\| \leq G^2$. $(b3)$ is based on the assumption $\frac{\eta_{t_0}}{\eta_t} \leq 2$.

Proof of Lemma 4. If we define $w_t = (a+t)^2$ and $\eta_t = \frac{4}{\mu(a+t)}$, we have:

$$a_{t+1} \leq (1-\mu\eta_t)a_t - A\eta_t e_t + \eta_t^2 B - \eta_t^3 C. \tag{A.13}$$

Now multiply equation A.13 with $\frac{w_t}{\eta_t}$, which yields: $a_{t+1}\frac{w_t}{\eta_t} \leq (1-\mu\eta_t)\frac{w_t}{\eta_t}a_t - w_t e_t A + w_t\eta_t B - w_t\eta_t^2 C$. By recursively replacing $\frac{w_{t-1}}{\eta_{t-1}}$, we obtain: $a_T\frac{w_{T-1}}{\eta_{T-1}} \leq (1-\mu\eta_0)\frac{w_0}{\eta_0}a_0 - \sum_{t=0}^{T-1}w_t e_t A + \sum_{t=0}^{T-1}w_t\eta_t B - \sum_{t=0}^{T-1}w_t\eta_t^2 C$. i.e.

$$A\sum_{t=0}^{T-1}w_t e_t \leq (1-\mu\eta_0)\frac{w_0}{\eta_0}a_0 + \sum_{t=0}^{T-1}w_t\eta_t B - \sum_{t=0}^{T-1}w_t\eta_t^2 C. \tag{A.14}$$

We will now derive upper bounds for the terms on the right hand side. We have $\frac{w_0}{\eta_0} = \frac{\mu a^3}{4}$. Then:

$$\sum_{t=0}^{T-1}w_t\eta_t = \sum_{t=0}^{T-1}\frac{4(a+t)}{\mu} = \frac{2T^2+4aT-2T}{\mu} \leq \frac{2T(T+2a)}{\mu},\ \sum_{t=0}^{T-1}w_t\eta_t^2 = \sum_{t=0}^{T-1}\frac{16}{\mu^2} = \frac{16T}{\mu^2}. \tag{A.15}$$

Let $S_T := \sum_{t=0}^{T-1}w_t = \sum_{t=0}^{T-1}(a+t)^2 = \frac{T}{6}(2T^2+6aT-3T+6a^2-6a+1)$. Then $S_T \geq \frac{T^3}{3}+aT^2-\frac{T^2}{2}+a^2T-aT \overset{(c1)}{\geq}\frac{T^3}{3}$, $(c1)$ is due to $aT^2-\frac{T^2}{2}+a^2T-aT = T^2(a-\frac{1}{2})+aT(a-1) \geq 0$, where $a \geq 1$ and $T \geq 0$.

Taking into account Assumption 1, Assumption 3, and the AFLAM convergence proof process, we can provide the parameter value ranges as follows: In Lemma 1, it is required that $\eta_t \leq \frac{1}{4L}$. In Lemma 3, the condition is that: $\eta_t = \frac{4}{\mu(a+t)}$ and $\frac{\eta_{t0}}{\eta_t} \leq 2$, so $\eta_t = \frac{4}{\mu(a+t)} \leq \frac{1}{4L}$, which implies $a \geq \frac{16L}{\mu}$. Also, $\frac{\eta_{t0}}{\eta_t} \leq 2$,which means $\frac{\frac{4}{\mu a}}{\frac{4}{\mu(a+t)}} \leq 2$, leading to $a \geq E$. Therefore $a = max\{\frac{16L}{\mu}, E\}$.

A.2. Proof of Theorem

Proof of Theorem 1. From Lemma 1, we can deduce that

$$\begin{aligned}\mathbb{E}\left\|\overline{\theta}_{t+1}-\theta^*\right\|^2 \leq (1-\mu\eta_t)\mathbb{E}\left\|\overline{\theta}_t-\theta^*\right\|^2 + \underbrace{2(1-L\eta_t)\sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|\overline{\theta}_t-\theta_{t,k}\right\|^2}_{F_1} \\ -\frac{3}{4}\eta_t\mathbb{E}\left[\ell(\overline{\theta}_t)-\ell(\theta^*)\right] + \mathbb{E}\left[6L\eta_t^2\mathcal{Q}-4L^2\eta_t^3\mathcal{Q}\right] + \underbrace{\eta_t^2\mathbb{E}\left\|\overline{g}_t-g_t\right\|^2}_{F_2} + \underbrace{2\eta_t\mathbb{E}\left\langle\overline{\theta}_t-\theta^*-\eta_t\overline{g}_t,\,\overline{g}_t-g_t\right\rangle}_{F_3}.\end{aligned} \tag{A.16}$$

Let's calculate the bounds for each term separately. First, we can obtain the bound for the term F_1 from Lemma 3:

$$F_1 = 2(1-L\eta_t)\sum_{k=1}^{K}\varphi_{t,k}\mathbb{E}\left\|\overline{\theta}_t-\theta_{t,k}\right\|^2 \leq 8\eta_t^2(1-L\eta_t)E^2G^2. \tag{A.17}$$

Secondly, we can get the bound for the term F_2 from Lemma 2: $F_2 = \eta_t^2\mathbb{E}\left\|\overline{g}_t-g_t\right\|^2 \leq \eta_t^2\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2$. Finally, we can derive the bound for the term F_3 from $\mathbb{E}\left\|\overline{g}_t\right\| = \mathbb{E}\left\|g_t\right\|$: $F_3 = 2\eta_t\mathbb{E}\langle\overline{\theta}_t-\theta^*-\eta_t\overline{g}_t,\,\overline{g}_t-g_t\rangle = 0$.

In conclusion, when $\eta_t \leq \frac{1}{4L}$, η_t is non-increasing, and $a = max\{\frac{16L}{\mu}, E\}$, we have:

$$\begin{aligned}&\mathbb{E}\left\|\overline{\theta}_{t+1}-\theta^*\right\|^2 \leq (1-\mu\eta_t)\mathbb{E}\left\|\overline{\theta}_t-\theta^*\right\|^2 + 8\eta_t^2(1-L\eta_t)E^2G^2 - \frac{3}{4}\eta_t\mathbb{E}\left[\ell(\overline{\theta}_t)-\ell(\theta^*)\right] \\ &+\eta_t^2\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2 + 6L\eta_t^2\mathcal{Q} - 4L^2\eta_t^3\mathcal{Q} \\ &\leq (1-\mu\eta_t)\,\mathbb{E}\left\|\overline{\theta}_t-\theta^*\right\|^2 - \frac{3}{4}\eta_t\mathbb{E}\left[\ell(\overline{\theta}_t)-\ell(\theta^*)\right] + \left(8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}\right)\eta_t^2 - \left(8LE^2G^2+4L^2\mathcal{Q}\right)\eta_t^3.\end{aligned}$$

As known from Lemma 4:

$$A\sum_{t=0}^{T-1}w_te_t \leq (1-\mu\eta_0)\frac{w_0}{\eta_0}a_0 + \sum_{t=0}^{T-1}w_t\eta_tB - \sum_{t=0}^{T-1}w_t\eta_t^2C \leq \frac{w_0}{\eta_0}a_0 + \frac{2T(T+2a)}{\mu}B - \frac{16T}{\mu^2}C. \tag{A.18}$$

Now multiply equation (A.18) with $\frac{1}{S_T}$, which yields:

$$\frac{A}{S_T}\sum_{t=0}^{T-1}w_te_t \leq \frac{w_0}{\eta_0S_T}a_0 + \frac{2T(T+2a)}{\mu S_T}B - \frac{16T}{\mu^2S_T}C, \tag{A.19}$$

for $A=\frac{3}{4}$, $B=8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}$, $C=(8LE^2G^2+4L^2\mathcal{Q})$, $\mathcal{Q}=\ell(\theta^*)-\sum_{k=1}^{K}\varphi_{t,k}\ell_k^*$, $e_t = \mathbb{E}\left[\ell(\overline{\theta}_t)-\ell(\theta^*)\right]$. We know that $S_T \geq \frac{T^3}{3}$ and due to the property of the convex function $\ell(\cdot)$, we have: $\mathbb{E}\left[\ell(\hat{\theta}_T)-\ell(\theta^*)\right] \leq \frac{1}{S_T}\sum_{t=0}^{T-1}w_t\mathbb{E}\left[\ell(\overline{\theta}_t)-\ell(\theta^*)\right]$, where $\hat{\theta}_T = \frac{1}{S_T}\sum_{k=1}^{K}\sum_{t=0}^{T-1}w_t\varphi_{t,k}\theta_{t,k}$,$w_t=(a+t)^2$,$\eta_t = \frac{4}{\mu(a+t)}$,$\frac{w_0}{\eta_0}=\frac{\mu a^3}{4}$.

$$\mathbb{E}\left[\ell(\hat{\theta}_T)-\ell(\theta^*)\right] \leq \frac{\mu a^3}{3S_T}||\overline{\theta}_0-\theta^*||^2 + \frac{8T(T+2a)}{3\mu S_T}\left(8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}\right) - \frac{64T}{3\mu^2S_T}(8LE^2G^2+4L^2\mathcal{Q}).$$

We utilize Lemma 2 from Rakhlin et al. (2012): $\mathbb{E}\left\|\theta_0-\theta^*\right\|^2 \leq \frac{4G^2}{\mu^2}$ for u-strongly convex ℓ. When both $\hat{\theta}_T$ and a meet the conditions mentioned earlier, we can derive the following results:

$$\mathbb{E}\left[\ell(\hat{\theta}_T)-\ell(\theta^*)\right] \leq \frac{4G^2a^3}{3\mu S_T} + \frac{8T(T+2a)}{3\mu S_T}\left(8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}\right) - \frac{64T}{3\mu^2S_T}\left(8LE^2G^2+4L^2\mathcal{Q}\right).$$

Appendix B: Convergence Proof for Partial Client Participation Under Convex Condition

B.1. Proof of Lemmas

Proof of Lemma 10. Let $\{a_i\}_{i=1}^K$ represent any fixed deterministic sequence. For each value, the probability is denoted as p_i'. We draw M values to form the set S_t, where $S_t = \{i_1, i_2, \ldots, i_M\} \subset [K]$. Then,

$$\mathbb{E}_{S_t} a_i = \mathbb{E}_{S_t} \sum_{k=1}^{M} a_{i_k} = M\mathbb{E}_{S_t} a_{i1} = M \sum_{k=1}^{K} p_k^{'} x_k. \tag{B.1}$$

Given Equation B.1 holds for both uniform and non-uniform client sampling (i.e., whether $p_i^{'}$ is equal or not).

Proof of Lemma 11. We have $\grave{\theta}_t = \sum_{k\in S_t} \frac{K}{M}\varphi_{t,k}\theta_{t,k}$, where the balancing factors are bounded: $\tilde{\epsilon} \le \varphi_{t,k} \le \zeta$.

$$\begin{aligned}
&\mathbb{E}_{S_t}\left\|\grave{\theta}_t - \tilde{\theta}_t\right\|^2 = \mathbb{E}_{S_t}\left\|\sum_{k\in S_t}\frac{K}{M}\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2 = \frac{1}{M^2}\mathbb{E}_{S_t}\left\|\sum_{k=1}^{K}\mathbb{I}\{k\in S_t\}\left(K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right)\right\|^2 \\
&= \frac{1}{M^2}\left[\sum_{k=1}^{K}\mathbb{P}\{k\in S_t\}\left\|K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2 + \sum_{k_i\neq k_j}\mathbb{P}(k_i,k_j\in S_t)\left\langle K\varphi_{t,k_i}\theta_{t,k_i} - \tilde{\theta}_t, K\varphi_{t,k_j}\theta_{t,k_j} - \tilde{\theta}_t\right\rangle\right] \\
&\overset{(c1)}{\le} \frac{\mathcal{G}}{M^2}\left[\sum_{k=1}^{K}\left\|K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2 + \overline{\mathcal{F}}\sum_{k_i\neq k_j}\left\langle K\varphi_{t,k_i}\theta_{t,k_i} - \tilde{\theta}_t, K\varphi_{t,k_j}\theta_{t,k_j} - \tilde{\theta}_t\right\rangle\right] \overset{(c2)}{=} \frac{\mathcal{G}\mathcal{F}}{M^2}\underbrace{\sum_{k=1}^{K}\left\|K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2}_{G}.
\end{aligned}$$

$(c1)$ is due to the fact that $\mathbb{P}\{k\in S_t\} = 1 - \prod_{v=1}^{|S_t|}\left(1 - \frac{p_k}{\sum_{k'\in\mathcal{K}_v} p_{k'}}\right)$. Here, $\mathcal{K}_v$ represents the set of remaining elements during the v-th sampling. Since client sampling in FL is performed without replacement, the probability that element k is ultimately selected into the subset S_t is calculated. $\mathbb{P}\{k_i,k_j\in S_t\} = \mathbb{P}\{k_i\in S_t\}\left(1 - \prod_{v=1}^{|S_t|-1}\left(1-p_k^{'}\right)\right)$, where, $p_k^{'} = \frac{p_k}{\sum_{k'\in\mathcal{K}_{(v,i)}} p_{k'}}$,$\mathcal{K}_{(v,i)}$ denotes the set of remaining elements during the v-th sampling after excluding client i. We introduce the upper bound of the probability $\overline{p_k} = \max([p_{k,k\in[K]}])$ and the lower bound $\underline{p_k} = \min([p_{k,k\in[K]}])$. Then, $\mathbb{P}\{k\in S_t\} \le 1 - \prod_{v=1}^{|S_t|}\left(1 - \frac{\overline{p_k}}{\underline{p_k}\cdot|\mathcal{K}_v|}\right) = \mathcal{G}$, $\mathbb{P}\{k_i,k_j\in S_t\} = \mathbb{P}\{k_i\in S_t\}\left(1 - \prod_{v=1}^{|S_t|-1}\left(1-p_k^{'}\right)\right) \le \mathbb{P}\{k_i\in S_t\}\left(1 - \prod_{v=1}^{|S_t|-1}\left(1 - \frac{\overline{p_k}}{\underline{p_k}\cdot|\mathcal{K}_{v,i}|}\right)\right) = \overline{\mathcal{F}}$. let $\mathcal{F} = \prod_{v=1}^{|S_t|-1}\left(1 - \frac{\overline{p_k}}{\underline{p_k}\cdot|\mathcal{K}_{v,i}|}\right)$. $(c2)$ uses the equality $\sum_{k=1}^{K}\left\|K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2 + \sum_{k_i\neq k_j}\left\langle K\varphi_{t,k_i}\theta_{t,k_i} - \tilde{\theta}_t, K\varphi_{t,k_j}\theta_{t,k_j} - \tilde{\theta}_t\right\rangle = 0$.

Next, we will prove the upper bound for the term G. The proof for G is similar to Lemma 3 in the convergence analysis when all clients participate. Therefore, we'll streamline the identical steps during the proof.

$$\begin{aligned}
G &= \sum_{k=1}^{K}\left\|K\varphi_{t,k}\theta_{t,k} - \tilde{\theta}_t\right\|^2 = \sum_{k=1}^{K}\left\|\left(K\varphi_{t,k}\theta_{t,k} - K\varphi_{t,k}\overline{\theta}_{t0}\right) - \left(\tilde{\theta}_t - K\varphi_{t,k}\overline{\theta}_{t0}\right)\right\|^2 \\
&\le \sum_{k=1}^{K}\left\|K\varphi_{t,k}\theta_{t,k} - K\varphi_{t,k}\overline{\theta}_{t0}\right\|^2 = K^2\sum_{k=1}^{K}\varphi_{t,k}^2\left\|\theta_{t,k} - \overline{\theta}_{t0}\right\|^2 \le K^2\sum_{k=1}^{K}\varphi_{t,k}\left\|\theta_{t,k} - \overline{\theta}_{t0}\right\|^2 \le 4\eta_t^2K^2E^2G^2.
\end{aligned} \tag{B.2}$$

The final inequality is proven in Lemma 3's proof, so get: $\mathbb{E}_{S_t}\left\|\grave{\theta}_t - \tilde{\theta}_t\right\|^2 \le \frac{\mathcal{G}\mathcal{F}}{M^2}4\eta_t^2K^2E^2G^2$.

B.2. Proof of Theorem

Proof of theorem 3. $\mathbb{E}\left\|\grave{\theta}_{t+1} - \theta^*\right\|^2 = \underbrace{\mathbb{E}\left\|\grave{\theta}_{t+1} - \tilde{\theta}_{t+1}\right\|^2}_{H_1} + \underbrace{\mathbb{E}\left\|\tilde{\theta}_{t+1} - \theta^*\right\|^2}_{H_2} + \underbrace{2\mathbb{E}\left\langle\grave{\theta}_{t+1} - \tilde{\theta}_{t+1}, \tilde{\theta}_{t+1} - \theta^*\right\rangle}_{H_3}$.

Next, we will derive the bounds for H_1, H_2, and H_3 separately. First, from Lemma 10, we can deduce that $H_3 = 0$. Second, according to Lemma 11, we have $H_1 \le \frac{\mathcal{G}\mathcal{F}}{M^2}4K^2\eta_t^2E^2G^2$. Finally, H_2 corresponds to the result of Lemma 1 under full client participation. Therefore, under partial client participation, $\mathbb{E}\left\|\grave{\theta}_{t+1} - \theta^*\right\|^2$ satisfies:

$$\mathbb{E}\left\|\dot{\theta}_{t+1}-\theta^*\right\|^2 \leq (1-\mu\eta_t)\mathbb{E}\left\|\tilde{\theta}_t-\theta^*\right\|^2 - \frac{3}{4}\eta_t\mathbb{E}\left[\ell(\tilde{\theta}_t)-\ell(\theta^*)\right] + \left(8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}+\frac{\mathcal{GF}}{M^2}4K^2E^2G^2\right)\eta_t^2 - \left(8LE^2G^2+4L^2\mathcal{Q}\right)\eta_t^3. \tag{B.3}$$

Let $\dot{\theta}_T=\frac{1}{S_T}\sum_{t=0}^{T-1}\sum_{k\in S_t}w_t\frac{K}{M}\varphi_{t,k}\theta_{t,k}$, and $\mathcal{Q}=\ell(\theta^*)-\sum_{k=1}^{K}\varphi_{t,k}\ell_k^*$, then equation (B.4) holds.

$$\mathbb{E}\left[\ell(\dot{\theta}_T)-\ell(\theta^*)\right] \leq \frac{4a^3G^2}{3\mu S_T}+\frac{8T(T+2a)}{3\mu S_T}\left(8E^2G^2+\sum_{k=1}^{K}\varphi_{t,k}^2\sigma_k^2+6L\mathcal{Q}+\frac{\mathcal{GF}}{M^2}4K^2E^2G^2\right) - \frac{64T}{3\mu^2S_T}\left(8LE^2G^2+4L^2\mathcal{Q}\right), \tag{B.4}$$

It's important to note that in this context, $\mathbb{E}\left\|\dot{\theta}_{t+1}-\theta^*\right\|^2$ involves $\dot{\theta}_{t+1}=\sum_{k\in S_{t+1}}\frac{K}{M}\varphi_{t+1,k}\theta_{t+1,k}$, which is not the same as the algorithm $\bar{\theta}_{t+1}=\sum_{k\in S_{t+1}}\frac{\varphi_{t+1,k}}{\sum_{k'\in S_{t+1}}\varphi_{t+1,k}}\theta_{t+1,k'}$ designed in Section 3. Therefore, we need to further modify the local loss functions. Let $\vartheta=\sum_{k\in S_t}\varphi_{t,k}$ denote the sum of balance factors of the clients sampled in the t-th iteration. In Section 3, we know that $\varphi_{t,k}$ has upper and lower bounds. Let's denote them as $\tilde{\epsilon}\leq\varphi_{t,k}\leq\zeta$. Consequently, we have $M\tilde{\epsilon}\leq\vartheta\leq 1$. When $M=K$, $\vartheta=1$. Now, let $\tilde{\ell}_k(\theta)=\frac{M}{K\vartheta}\ell_k(\theta)$. This operation effectively scales the local loss functions of the clients. The global objective then becomes $\min_{\theta\in\mathbb{R}^d}\ell(\theta)=\sum_{k=1}^{K}\varphi_{t,k}\tilde{\ell}_k(\theta)$. It can be observed that (B.4) still holds in this case. Here, we define $\tilde{L}\triangleq\omega L$, $\tilde{\mu}\triangleq\upsilon\mu$, $\tilde{\sigma}_k\triangleq\sqrt{\omega}\sigma$, and $\tilde{G}\triangleq\omega G$, where $\omega=\frac{M}{K}max\left\{\frac{1}{\sum_{k\in S_t}\varphi_{t,k}}\right\}$ and $\upsilon=\frac{M}{K}min\left\{\frac{1}{\sum_{k\in S_t}\varphi_{t,k}}\right\}$. Then, for $\hat{\theta}_T=\frac{1}{S_T}\sum_{t=0}^{T-1}\sum_{k\in S_t}w_t\frac{\varphi_{t,k}}{\vartheta}\theta_{t,k}$, the following equation holds:

$$\mathbb{E}\left[\ell(\hat{\theta}_T)-\ell(\theta^*)\right] \leq \frac{4a^3\tilde{G}^2}{3\tilde{\mu}S_T}+\frac{8T(T+2a)}{3\tilde{\mu}S_T}\left(8E^2\tilde{G}^2+\sum_{k=1}^{K}\varphi_{t,k}^2\tilde{\sigma}_k^2+6\tilde{L}\mathcal{Q}+\frac{\mathcal{GF}}{M^2}4K^2E^2\tilde{G}^2\right) - \frac{64T}{3\tilde{\mu}^2S_T}\left(8\tilde{L}E^2\tilde{G}^2+4\tilde{L}^2\mathcal{Q}\right).$$

Appendix C: Convergence Proof for Full Client Participation Under Non-convex Condition

C.1. Proof of Lemmas

Proof of Lemma 5. The goal is to compute the upper bound of $\mathbb{E}\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\|^2$. First, we have:

$$\mathbb{E}\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\|^2=\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2+\mathbb{E}\left\|\nabla\ell_k(\overline{\theta}_t)\right\|^2-2\mathbb{E}\left\langle\nabla\ell_k(\theta_{t,k},\xi_t^{k,i}),\nabla\ell_k(\overline{\theta}_t)\right\rangle. \tag{C.1}$$

According to Assumption 4 (the gradient upper bound assumption), we have $\mathbb{E}\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\|^2\leq G^2$. Then,

$$\mathbb{E}\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\|^2\leq G^2+\mathbb{E}\left\|\nabla\ell_k(\overline{\theta}_t)\right\|^2-2\mathbb{E}\left\langle\nabla\ell_k(\theta_{t,k},\xi_t^{k,i}),\nabla\ell_k(\overline{\theta}_t)\right\rangle. \tag{C.2}$$

To simplify the inner product term, we introduce the variable: $\mathcal{Z}=\sqrt{\mathbb{E}\|\nabla\ell_k(\overline{\theta}_t)\|^2}$. Using the Cauchy-Schwarz inequality, we have: $\mathbb{E}\left\langle\nabla\ell_k(\theta_{t,k}\xi_t^{k,i}),\nabla\ell_k(\overline{\theta}_t)\right\rangle\leq\sqrt{\mathcal{Z}^2\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2}$. Substituting this and simplifying:

$$G^2+\mathcal{Z}^2-2\sqrt{\mathcal{Z}^2\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2}\leq\sigma_k^2. \tag{C.3}$$

According to Assumption 3, we have $\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\theta_{t,k})\right\|^2\leq\sigma_k^2$. So,$(G-Z)^2\leq\sigma_k^2$ $(when\ G\geq 0)$, $(G+Z)^2\leq\sigma_k^2$ $(when\ G<0)$. We have $(G-|\sigma_k|)\leq Z\leq(G+|\sigma_k|)$, so $Z^2\leq(|G|+|\sigma_k|)^2$, i.e., $\mathbb{E}\|\nabla\ell_k(\overline{\theta}_t)\|^2\leq(|G|+|\sigma_k|)^2$ is have upper bound.

Proof of Lemma 6. Using the notation E and t_0 from the proof of Lemma 3, we further prove this Lemma.

$$\begin{aligned}
&\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\right\|^2=\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_{t_0})-(\nabla\ell_k(\overline{\theta}_t)-\ell_k(\overline{\theta}_{t_0}))\right\|^2\\
&\leq\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_{t_0})\right\|^2-\left\|\mathbb{E}[\nabla\ell_k(\overline{\theta}_t)-\ell_k(\overline{\theta}_{t_0})]\right\|^2\overset{(d1)}{\leq}\mathbb{E}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_{t_0})\right\|^2\\
&\overset{(d2)}{\leq}L^2\mathbb{E}\left\|\theta_{t,k}-\overline{\theta}_{t_0}\right\|^2\leq L^2\mathbb{E}\sum_{t=t_0}^{t_0+E-1}\eta_t^2\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2\leq 4L^2\eta_t^2E^2G^2,
\end{aligned}\tag{C.4}$$

$d1$ is due to $\mathbb{E}\|X-\mathbb{E}X\|^2=\mathbb{E}\|X\|^2-\|\mathbb{E}X\|^2$. $(d2)$ is the property of L-smoothness.

Proof of Lemma 7. The goal is to compute the upper bound of $\ell(\overline{\theta}_{t+1})-\ell(\overline{\theta}_t)$.

$$\ell(\overline{\theta}_{t+1})-\ell(\overline{\theta}_t)\leq\left\langle\nabla\ell(\overline{\theta}_t),\overline{\theta}_{t+1}-\overline{\theta}_t\right\rangle+\frac{L}{2}\left\|\overline{\theta}_{t+1}-\overline{\theta}_t\right\|^2\leq\underbrace{-\eta_t\left\langle g_t,\nabla\ell(\overline{\theta}_t)\right\rangle}_{H}+\frac{L\eta_t^2}{2}\left\|g_t\right\|^2.\tag{C.5}$$

$$\begin{aligned}
H&=-\eta_t\left\langle g_t,\nabla\ell(\overline{\theta}_t)\right\rangle=-\frac{\eta_t}{2}\left[\|g_t\|^2+\|\nabla\ell(\overline{\theta}_t)\|^2-\|g_t-\nabla\ell(\overline{\theta}_t)\|^2\right]\\
&=-\frac{\eta_t}{2}\|g_t\|^2-\frac{\eta_t}{2}\|\nabla\ell(\overline{\theta}_t)\|^2+\frac{\eta_t}{2}\|g_t-\nabla\ell(\overline{\theta}_t)\|^2\leq-\frac{\eta_t}{2}\|g_t\|^2-\mu\eta_t(\ell(\overline{\theta}_t)-\ell(\theta^*))+\frac{\eta_t}{2}\underbrace{\|g_t-\nabla\ell(\overline{\theta}_t)\|^2}_{H_1}.
\end{aligned}\tag{C.6}$$

The final inequality follows from the μ-PL in Assumption 2: $\|\nabla\ell(\overline{\theta}_t)\|^2\geq 2\mu(\ell(\overline{\theta}_t)-\ell(\theta^*))$.

$$\begin{aligned}
H_1&=\|g_t-\nabla\ell(\overline{\theta}_t)\|^2=\left\|\sum_{k=1}^K\varphi_{t,k}\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell(\overline{\theta}_t)\right\|^2=\left\|\sum_{k=1}^K\varphi_{t,k}\left[\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\right]\right\|^2\\
&\leq K\sum_{k=1}^K\varphi_{t,k}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\right\|^2.
\end{aligned}\tag{C.7}$$

The final inequality follows from the bound $||\sum_{k=1}^K a_t||^2\leq K\sum_{k=1}^K||a_t||^2$.

$$H\leq-\frac{\eta_t}{2}\|g_t\|^2-\mu\eta_t(\ell(\overline{\theta}_t)-\ell(\theta^*))+\frac{K\eta_t}{2}\sum_{k=1}^K\varphi_{t,k}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\right\|^2.\tag{C.8}$$

Substituting H into Equation (C.5), we get:

$$\ell(\overline{\theta}_{t+1})-\ell(\overline{\theta}_t)\leq-\mu\eta_t[\ell(\overline{\theta}_t)-\ell(\theta^*)]+\frac{K\eta_t}{2}\sum_{k=1}^K\varphi_{t,k}\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})-\nabla\ell_k(\overline{\theta}_t)\right\|^2+\frac{\eta_t(L\eta_t-1)}{2}\left\|g_t\right\|^2.\tag{C.9}$$

Proof of Lemma 8. In proving the upper bound of $\mathbb{E}\left\|g_t\right\|^2$, we consider the impact of heterogeneity on the results based on the client data distribution, analyzing both scenarios with and without heterogeneity. When the heterogeneity of client data is not considered, then

$$\mathbb{E}\left\|g_t\right\|^2=\mathbb{E}\left\|\sum_{k=1}^K\varphi_{t,k}\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2\overset{(d)}{\leq}K\sum_{k=1}^K\varphi_{t,k}^2\mathbb{E}\left\|\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2\leq K\sum_{k=1}^K\varphi_{t,k}^2G^2,\tag{C.10}$$

(d) includes the inequality $||\sum_{k=1}^K a_t||^2\leq K\sum_{k=1}^K||a_t||^2$, and the final step is due to the Assumption 4.

When the heterogeneity of client data is considered, compute the upper bound of $\mathbb{E}\left\|g_t\right\|^2$.

$$\mathbb{E}\left\|g_t\right\|^2=\mathbb{E}\left\|\sum_{k=1}^K\varphi_{t,k}\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2\leq K\mathbb{E}\sum_{k=1}^K\varphi_{t,k}^2\left\|\nabla\ell_k(\theta_{t,k},\xi_t^{k,i})\right\|^2\leq 2KL\mathbb{E}\underbrace{\left[\sum_{k=1}^K\varphi_{t,k}\ell_k(\theta_{t,k},\xi_t^{k,i})-\ell_k^*\right]}_{I}.$$

Then, to further bound the term I.

$$
\begin{aligned}
I &= \sum_{k=1}^{K} \varphi_{t,k} \ell_k(\theta_{t,k}, \xi_t^{k,i}) - \ell_k^* = \sum_{k=1}^{K} \varphi_{t,k} \left[\ell_k(\theta_{t,k}, \xi_t^{k,i}) - \ell_k(\overline{\theta}_t) \right] + \sum_{k=1}^{K} \varphi_{t,k} \left[\ell_k(\overline{\theta}_t) - \ell_k^* \right] \\
&\leq \sum_{k=1}^{K} \varphi_{t,k} \left\langle \nabla \ell_k(\overline{\theta}_t), \theta_{t,k} - \overline{\theta}_t \right\rangle + \frac{L}{2} \sum_{k=1}^{K} \varphi_{t,k} \|\theta_{t,k} - \overline{\theta}_t\|^2 + \sum_{k=1}^{K} \varphi_{t,k} \left[\ell_k(\overline{\theta}_t) - \ell_k^* \right] \\
&\leq \frac{\eta_t}{2} \sum_{k=1}^{K} \varphi_{t,k} \|\nabla \ell_k(\overline{\theta}_t)\|^2 + \frac{1}{2} (\frac{1}{\eta_t} + L) \sum_{k=1}^{K} \varphi_{t,k} \|\theta_{t,k} - \overline{\theta}_t\|^2 + \left[\ell(\overline{\theta}_t) - \sum_{k=1}^{K} \varphi_{t,k} \ell_k^* \right].
\end{aligned} \tag{C.11}
$$

The final step is the inequality $2\langle a, b \rangle \leq \lambda ||a||^2 + \frac{1}{\lambda} ||b||^2, \lambda = \eta_t > 0$. By combining I, we have:

$$
\mathbb{E} \left\| g_t \right\|^2 \leq KL\eta_t \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \|\nabla \ell_k(\overline{\theta}_t)\|^2 + KL(\frac{1}{\eta_t} + L) \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \|\theta_{t,k} - \overline{\theta}_t\|^2 + 2KL \left[\ell(\overline{\theta}_t) - \sum_{k=1}^{K} \varphi_{t,k} \ell_k^* \right]. \tag{C.12}
$$

Proof of Lemma 9. Following the proof strategy of Lemma 4, let $a_{t+1} \leq (1 - \mu\eta_t) a_t + \eta_t^2 A + \eta_t^3 B$, then

$$
a_T \leq (1 - \mu\eta_0) \frac{a^3}{(a+T)^3} a_0 + \frac{2T(T+2a)}{\mu^2 (a+T)^3} A + \frac{64T}{\mu^3 (a+T)^3} B. \tag{C.13}
$$

Similarly, by transforming $a_{t+1} \leq (1 - \mu\eta_t + KL\eta_t(L\eta_t - 1)) a_t + \eta_t^2 A + \eta_t^3 B + \eta^4 C$, we obtain:

$$
a_T \leq (1 - \mu\eta_0 + KL\eta_0(L\eta_0 - 1)) \frac{a^3}{(a+T)^3} a_0 + \frac{8T(T+2a)}{\mu^2 (a+T)^3} A + \frac{64T}{\mu^3 (a+T)^3} B + \frac{16T}{\mu^4 (a+T)^3} C. \tag{C.14}
$$

In this process, we use $\sum_{t=0}^{T-1} w_t \eta_t^3 = \sum_{t=0}^{T-1} \frac{4}{\mu^3 (a+T)} = \frac{4T}{\mu^3 (a+T)}$.

C.2. Proof of Theorem

Proof of theorem 2. Under non-convex conditions, without considering Non-IID, according to Lemma 7, we have

$$
\begin{aligned}
\mathbb{E}[\ell(\overline{\theta}_{t+1}) - \ell(\theta^*)] &\leq (1 - \mu\eta_t) \mathbb{E}[\ell(\overline{\theta}_t) - \ell(\theta^*)] + \frac{K\eta_t}{2} \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \left\| \nabla \ell_k(\theta_{t,k}, \xi_t^{k,i}) - \nabla \ell_k(\overline{\theta}_t) \right\|^2 + \frac{\eta_t (L\eta_t - 1)}{2} \mathbb{E} \left\| g_t \right\|^2 \\
&\leq (1 - \mu\eta_t) \mathbb{E}[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 2L^2 K \eta_t^3 E^2 G^2 + \frac{K\eta_t (L\eta_t - 1)}{2} \sum_{k=1}^{K} \varphi_{t,k}^2 G^2 \\
&\leq (1 - \mu\eta_t) \mathbb{E}[\ell(\overline{\theta}_t) - \ell(\theta^*)] + \frac{KL\eta_t^2}{2} \sum_{k=1}^{K} \varphi_{t,k}^2 G^2 + 2L^2 K \eta_t^3 E^2 G^2.
\end{aligned} \tag{C.15}
$$

The penultimate inequality is based on Lemma 6 and Lemma 8. According to Lemma 9, we can obtain:

$$
\mathbb{E}[\ell(\overline{\theta}_{t+1}) - \ell(\theta^*)] \leq (1 - \mu\eta_0) \frac{a^3}{(a+T)^3} \mathbb{E}[\ell(\overline{\theta}_0) - \ell(\theta^*)] + \frac{T(T+2a)KL}{\mu^2 (a+T)^3} \sum_{k=1}^{K} \varphi_{t,k}^2 G^2 + \frac{128TL^2 K E^2 G^2}{\mu^3 (a+t)^3}. \tag{C.16}
$$

Under non-convex conditions, considering client data heterogeneity, according to Lemma 7, we have

$$
\begin{aligned}
&\mathbb{E}[\ell(\overline{\theta}_{t+1}) - \ell(\theta^*)] \leq (1 - \mu\eta_t)[\ell(\overline{\theta}_t) - \ell(\theta^*)] + \frac{K\eta_t}{2} \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \left\| \nabla \ell_k(\theta_{t,k}, \xi_t^{k,i}) - \nabla \ell_k(\overline{\theta}_t) \right\|^2 + \frac{\eta_t (L\eta_t - 1)}{2} \mathbb{E} \left\| g_t \right\|^2 \\
&\leq (1 - \mu\eta_t) \mathbb{E}[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 2L^2 K \eta_t^3 E^2 G^2 + \frac{\eta_t (L\eta_t - 1)}{2} \left[KL\eta_t \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \|\nabla \ell_k(\overline{\theta}_t)\|^2 \right] \\
&+ \frac{KL(L\eta_t - 1)}{2} \left[(1 + L\eta_t) \sum_{k=1}^{K} \varphi_{t,k} \mathbb{E} \|\theta_{t,k} - \overline{\theta}_t\|^2 \right] + KL\eta_t (L\eta_t - 1) \left[\ell(\overline{\theta}_t) - \sum_{k=1}^{K} \varphi_{t,k} \ell_k^* \right] \\
&\leq (1 - \mu\eta_t) \mathbb{E}[\ell(\overline{\theta}_t) - \ell(\theta^*)] + 2L^2 K \eta_t^3 E^2 G^2 + \frac{KL\eta_t^2 (L\eta_t - 1)}{2} (|G| + |\sigma_k|)^2 \\
&+ 2KL\eta_t^2 E^2 G^2 (L^2 \eta_t^2 - 1) + KL\eta_t (L\eta_t - 1) \left[\ell(\overline{\theta}_t) - \sum_{k=1}^{K} \varphi_{t,k} \ell_k^* \right]
\end{aligned}
$$

$$= (1-\mu\eta_t + KL\eta_t(L\eta_t-1))\mathbb{E}[\ell(\overline{\theta}_t)-\ell(\theta^*)] + KL\eta_t^2\left[\ell(\theta_*) - \sum_{k=1}^{K}\varphi_{t,k}\ell_k^* - (|G|+|\sigma_k|)^2 - 2E^2G^2\right] \\ + \eta_t^3(2KL^2E^2G^2 + \frac{KL^2}{2}(|G|+|\sigma_k|)^2) + 2K\eta_t^4L^3E^2G^2. \tag{C.17}$$

The final inequality is based on Lemmas 5, 6, and 8. We define $\mathcal{Q} = \ell(\theta^*) - \sum_{k=1}^{K}\varphi_{t,k}\ell_k^*$, and $\zeta = (|G|+|\sigma_k|)^2$. Then,

$$\mathbb{E}[\ell(\overline{\theta}_{t+1})-\ell(\theta^*)] \le (1-\mu\eta_t + KL\eta_t(L\eta_t-1))\mathbb{E}[\ell(\overline{\theta}_t)-\ell(\theta^*)] \\ + KL\eta_t^2(\mathcal{Q}-\zeta-2E^2G^2) + \eta_t^3(2KL^2E^2G^2 + \frac{KL^2}{2}\zeta) + 2K\eta_t^4L^3E^2G^2. \tag{C.18}$$

According to Lemma 9, we can obtain:

$$\mathbb{E}[\ell(\overline{\theta}_{t+1})-\ell(\theta^*)] \le (1-\mu\eta_0 + KL\eta_0(L\eta_0-1))\frac{a^3}{(a+T)^3}\mathbb{E}[\ell(\overline{\theta}_0)-\ell(\theta^*)] + \frac{8T(T+2a)}{\mu^2(a+T)^3}KL(\mathcal{Q}-\zeta-2E^2G^2) \\ + \frac{64T}{\mu^3(a+T)^3}(2KL^2E^2G^2 + \frac{KL^2}{2}\zeta) + \frac{32T}{\mu^4(a+T)^3}KL^3E^2G^2.$$

Appendix D: Convergence Proof for Partial Client Participation Under Non-convex Condition

In the proof of this section, we adopt the same definitions as in the convex function proof. That is, we have $\dot{\theta}_t = \sum_{k\in S_t}\frac{K}{M}\varphi_{t,k}\theta_{t,k}$, where the weighting factor is bounded, i.e., $\tilde{\epsilon} \le \varphi_{t,k} \le \zeta$. Therefore, even when ℓ and ℓ_k are non-convex functions, Lemma 10 and 11 still hold. Based on this conclusion, we further prove Lemma 12.

Proof of Lemma 12. $\mathbb{E}_{S_t}[\ell(\dot{\theta}_t) - \ell(\tilde{\theta}_t)] \le \mathbb{E}_{S_t}\left\langle \nabla\ell(\tilde{\theta}_t), \dot{\theta}_t - \tilde{\theta}_t\right\rangle + \frac{L}{2}\mathbb{E}_{S_t}\|\dot{\theta}_t - \tilde{\theta}_t\|^2$. From Lemma 10, we can deduce that the expectation of the first term is 0. Thus, based on Lemma 11, we have,

$$\mathbb{E}_{S_t}[\ell(\dot{\theta}_t) - \ell(\tilde{\theta}_t)] \le \frac{L}{2}\mathbb{E}_{S_t}\|\dot{\theta}_t - \tilde{\theta}_t\|^2 \le \frac{L\mathcal{GF}}{M^2}2\eta_t^2K^2E^2G^2.$$

Proof of Theorem 4. $\mathbb{E}[\ell(\dot{\theta}_t)-\ell(\theta^*)] = \mathbb{E}[\ell(\dot{\theta}_t)-\ell(\tilde{\theta}_t)] + \mathbb{E}[\ell(\tilde{\theta}_t)-\ell(\theta^*)] \le \frac{L\mathcal{GF}}{M^2}2\eta_t^2K^2E^2G^2 + \mathbb{E}[\ell(\tilde{\theta}_t)-\ell(\theta^*)]$. The last inequality is based on Lemma 12. So, according to Equation C.15, without considering heterogeneity, then:

$$\mathbb{E}[\ell(\dot{\theta}_t)-\ell(\theta^*)] \le (1-\mu\eta_t)\mathbb{E}[\ell(\overline{\theta}_t)-\ell(\theta^*)] + \eta_t^2G^2KL\left(\frac{1}{2}\sum_{k=1}^{K}\varphi_{t,k}^2 + \frac{\mathcal{GF}}{M^2}2KE^2\right) + 2L^2K\eta_t^3E^2G^2. \tag{D.1}$$

According to Lemma 9, we can obtain:

$$\mathbb{E}[\ell(\dot{\theta}_t)-\ell(\theta^*)] \le \frac{(1-\mu\eta_0)a^3}{(a+T)^3}\mathbb{E}[\ell(\overline{\theta}_0)-\ell(\theta^*)] + \frac{T(T+2a)KLG^2}{\mu^2(a+T)^3}\left(\sum_{k=1}^{K}\varphi_{t,k}^2 + \frac{\mathcal{GF}}{M^2}4KE^2\right) + \frac{128TL^2KE^2G^2}{\mu^3(a+T)^3}.$$

When considering heterogeneity, according to Equation C.18, then:

$$\mathbb{E}[\ell(\dot{\theta}_t)-\ell(\theta^*)] \le (1-\mu\eta_t + KL\eta_t(L\eta_t-1))\mathbb{E}[\ell(\overline{\theta}_t)-\ell(\theta^*)] + \eta_t^3(2KL^2E^2G^2 + \frac{KL^2}{2}\zeta) \\ + KL\eta_t^2(\mathcal{Q}-\zeta-2E^2G^2 + \frac{\mathcal{GF}}{M^2}2KE^2G^2) + 2K\eta_t^4L^3E^2G^2. \tag{D.2}$$

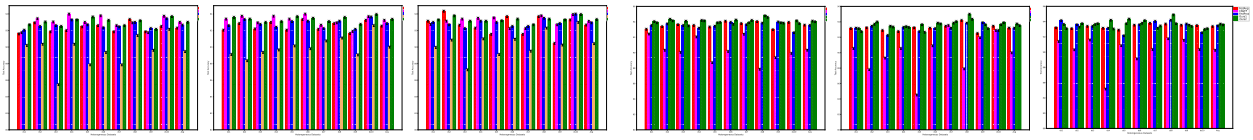
According to Lemma 9, we can obtain:

$$\mathbb{E}[\ell(\dot{\theta}_t)-\ell(\theta^*)] \le (1-\mu\eta_0 + KL\eta_0(L\eta_0-1))\frac{a^3}{(a+T)^3}\mathbb{E}[\ell(\overline{\theta}_0)-\ell(\theta^*)] + \frac{64T}{\mu^3(a+T)^3}(2KL^2E^2G^2 + \frac{KL^2}{2}\zeta) \\ + \frac{8T(T+2a)}{\mu^2(a+T)^3}KL(\mathcal{Q}-\zeta-2E^2G^2 + \frac{\mathcal{GF}}{M^2}2KE^2G^2) + \frac{32T}{\mu^4(a+T)^3}K\eta_t^4L^3E^2G^2.$$

Appendix E: AFLAM vs. baselines in complex models.

We compare the AFLAM with four SOTA baselines (FedAvg, FedProx, FedBN, CReFF) on USPS and MNIST using a more complex CNN model. In the experiments, we generated 10 Non-IID client datasets per dataset using a Dirichlet distribution and compared the results of each experiment with their average. Fig. E.1 (a) shows that with full participation, AFLAM outperforms three baselines and achieves comparable accuracy to CReFF. However, Fig. E.1 (b) reveals that AFLAM surpasses all baselines, including CReFF, under partial participation, where CReFF's performance declines. Fig. E.1 (b) also fully shows that the proposed scheme 1 and 2 can achieve the optimal results relative to other models under different E. By comparing the results of Fig. E.2 (a) and Fig. E.2 (b), it is clear that the AFLAM has the best performance on the MNIST, whether in the case of full participation (Fig. E.2 (a)) or partial participation (Fig. E.2 (b)).

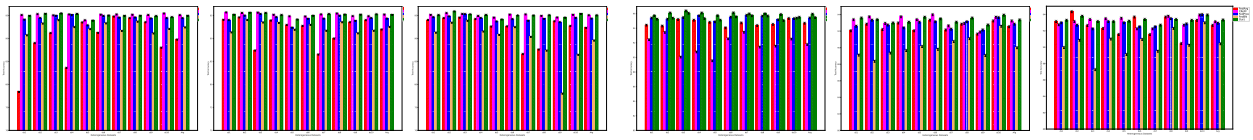
Figure E.1 Comparison results between AFLAM and baselines in the USPS dataset.



(a) Full clients participate (b) Partial clients participate

Note. In (a) and (b), E values (left to right): 1, 2, 5; $\mathrm{ds}_i\,(1 \leq i \leq 10)$ represents the i-th Non-IID client data distribution generated via Dirichlet sampling; Avg shows 10-experiment mean. In (b), Ours1 and Ours2 denote Schemes 1 and 2 of the AFLAM, respectively.

Figure E.2 Comparison results between AFLAM and baselines in the MNIST dataset.



(a) Full clients participate (b) Partial clients participate

Appendix F: Comparison between original data and attack reconstruction

Fig. F.1 compares attacker-reconstructed data via gradient inversion with original data.

Figure F.1 Comparison of Gradient Inversion Reconstruction Results.

(a) (b) (c)

References

Rakhlin A, Shamir O, Sridharan K (2012) Making gradient descent optimal for strongly convex stochastic optimization. *Proceedings of the 29th International Coference on International Conference on Machine Learning,* Madison, WI, USA, 1571–1578.