APPENDIX
PROOFS OF MAIN RESULTS

A. Proof of Theorem 1

Proof. The proof follows the discussion from Wang *et al.* [14]. Additionally, we consider clients' data heterogeneity as specified in Assumption 4, which is not considered by [14].

In this proof, we try to find the upper bound of $\min_{t\in[T]} \mathbb{E}\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2$, where $t \in [T]$ means $t \in \{0,\cdots,T\}$. First, we establish the inequality function (55) and find the upper bounds of T_1 and T_2, respectively. The term $\mathbb{E}\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2$ is derived from (79). Then, through practical assumptions on the learning rate η, we finally find the upper bound in equation (82).

First, we introduce some important notations. The locally averaged gradient $\boldsymbol{d}_n^{(t)}$ is normalized by a non-negative vector $\boldsymbol{a}_n \in \mathbb{R}^{r_n}$, with the definition as

$$\boldsymbol{d}_n^{(t)} = \frac{\boldsymbol{G}_n^{(t)}\boldsymbol{a}_n}{\|\boldsymbol{a}_n\|_1}, \tag{47}$$

where

$$\boldsymbol{G}_n^{(t)} = \left[g_n(\boldsymbol{\omega}_n^{(t,0)}), g_n(\boldsymbol{\omega}_n^{(t,1)}), \ldots, g_n(\boldsymbol{\omega}_n^{(t,r_n-1)})\right] \tag{48}$$

stacks all stochastic gradients in the t-th round.

$g_n(\boldsymbol{\omega}_n^{(t,k)})$ denotes the gradient of local loss function in the t-th round after k-local updates. $\boldsymbol{\omega}_n^{(t,k)}$ denotes client n's model after the k-th local update in the t-th communication round respectively. We let $\boldsymbol{a}_n(k)$ to denote the accumulation vector after performing k local steps on client n, then $\boldsymbol{a}_n(k) = [a_{n,0},\ldots,a_{n,k-1}]^\top$, where $a_{n,s}(k) \geq 0, \forall s \in \{0,\cdots,k-1\}$ is an arbitrary scalar. Thus, the local update rule is

$$\boldsymbol{\omega}_n^{(t,k)} - \boldsymbol{\omega}_n^{(t,0)} = -\eta\sum_{s=0}^{k-1} a_{n,s}(k) g_n(\boldsymbol{\omega}^{(t,s)}), \tag{49}$$

for any $k \geq 0$, where $a_{n,s}(k)$ is the s-th element in vector $\boldsymbol{a}_n(k) \in \mathbb{R}^k$. When $k = r_n$, we denote $\boldsymbol{a}_n = \boldsymbol{a}_n(r_n) = [a_{n,0},\ldots,a_{n,r_n-1}]^\top$.

Suppose the L1-norm of $\boldsymbol{a}_n$ is $a_n = ||\boldsymbol{a}_n||_1$, we define the following auxiliary variables

$$\boldsymbol{h}_n^{(t)} = \frac{1}{a_n}\sum_{k=0}^{r_n-1} a_{n,k}\nabla F_n(\boldsymbol{\omega}_n^{(t,k)}). \tag{50}$$

After matrix operation, we find that

$$\mathbb{E}[\boldsymbol{d}_n^{(t)} - \boldsymbol{h}_n^{(t)}] = 0. \tag{51}$$

In addition, since clients are independent of each other, we have

$$\mathbb{E}\langle \boldsymbol{d}_n^{(t)} - \boldsymbol{h}_n^{(t)}, \boldsymbol{d}_m^{(t)} - \boldsymbol{h}_m^{(t)}\rangle = 0, \forall n \neq m. \tag{52}$$

According to the L-smoothness Assumption 1 and the Descent Lemma [36], we have

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{\omega}^{(t+1,0)})] - F(\boldsymbol{\omega}^{(t,0)}) \leq \tfrac{L}{2}\mathbb{E}\left[\left\|\boldsymbol{\omega}^{(t+1,0)} - \boldsymbol{\omega}^{(t,0)}\right\|^2\right] \\ +\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{\omega}^{(t,0)}), \boldsymbol{\omega}^{(t+1,0)} - \boldsymbol{\omega}^{(t,0)}\right\rangle\right],\end{aligned} \tag{53}$$

where the expectation is taken over mini-batches $\xi_n^{(t,k)}, \forall n \in \{1,2,...,N\}, k \in \{0,1,...,r_n-1\}$ and $\langle a,b\rangle$ means the dot product of a and b.

Replacing the update rule in (1) into equation (53), we have

$$\begin{aligned}&\mathbb{E}[F(\boldsymbol{\omega}^{(t+1,0)})] - F(\boldsymbol{\omega}^{(t,0)}) \\ &\leq \frac{L}{2}\mathbb{E}\left[\left\|-(\sum_{n=1}^N \rho_n r_n)\sum_{n=1}^N \rho_n\eta\boldsymbol{d}_n^{(t)}\right\|^2\right] \\ &+\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{\omega}^{(t,0)}), -(\sum_{n=1}^N \rho_n r_n)\sum_{n=1}^N \rho_n\eta\boldsymbol{d}_n^{(t)}\right\rangle\right] \end{aligned} \tag{54}$$

$$\begin{aligned}&= \frac{(\sum_{n=1}^N \rho_n r_n)^2\eta^2 L}{2}\underbrace{\mathbb{E}\left[\left\|\sum_{n=1}^N \rho_n\boldsymbol{d}_n^{(t)}\right\|^2\right]}_{T_1} \\ &-(\sum_{n=1}^N \rho_n r_n)\eta\underbrace{\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{\omega}^{(t,0)}), \sum_{n=1}^N \rho_n\boldsymbol{d}_n^{(t)}\right\rangle\right]}_{T_2}.\end{aligned} \tag{55}$$

1) Bounding T_1 in equation (55):

$$T_1 = \mathbb{E}\left[\left\|\sum_{n=1}^N \rho_n\boldsymbol{d}_n^{(t)}\right\|^2\right] \tag{56}$$

$$\leq 2\mathbb{E}\left[\left\|\sum_{n=1}^N \rho_n(\boldsymbol{d}_n^{(t)} - \boldsymbol{h}_n^{(t)})\right\|^2\right] + 2\mathbb{E}\left[\left\|\sum_{n=1}^N \rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right] \tag{57}$$

$$= 2\sum_{n=1}^N \rho_n^2\mathbb{E}\left[\left\|\boldsymbol{d}_n^{(t)} - \boldsymbol{h}_n^{(t)}\right\|^2\right] + 2\mathbb{E}\left[\left\|\sum_{n=1}^N \rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right], \tag{58}$$

where (57) follows from the fact $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and (58) is adopted from equation (52).

Before we dive into equation (58), we adopt the following Lemma 6 [14].

Lemma 6. *Suppose $\{\boldsymbol{A}_k\}_{k=1}^T$ is a sequence of random matrices and $\mathbb{E}[\boldsymbol{A}_k|\boldsymbol{A}_{k-1},\boldsymbol{A}_{k-2},...,\boldsymbol{A}_1] = \boldsymbol{0}, \forall k$. Then,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^T \boldsymbol{A}_k\right\|_F^2\right] = \sum_{k=1}^T \mathbb{E}\left[\|\boldsymbol{A}_k\|_F^2\right], \tag{59}$$

where $\|\cdot\|_F$ means $F-$norm.

Replacing the definition of $\boldsymbol{d}_n^{(t)}$ in equation (47) and $\boldsymbol{h}_n^{(t)}$ in equation (50) into the first term in equation (58), it can be written as

$$
\begin{aligned}
&2\sum_{n=1}^{N}\rho_n^2\mathbb{E}\left[\left\|\boldsymbol{d}_n^{(t)}-\boldsymbol{h}_n^{(t)}\right\|^2\right]\\
&=2\sum_{n=1}^{N}\rho_n^2\mathbb{E}\left[\left\|\frac{\boldsymbol{G}_n^{(t)}\boldsymbol{a}_n}{\|\boldsymbol{a}_n\|_1}-\frac{1}{a_n}\sum_{k=0}^{r_n-1}a_{n,k}\nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\right\|^2\right] && (60)\\
&=2\sum_{n=1}^{N}\rho_n^2\mathbb{E}\left[\left\|\frac{1}{a_n}\sum_{k=0}^{r_n-1}a_{n,k}\big(g_n(\boldsymbol{\omega}_n^{(t,k)})-\nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\big)\right\|^2\right] && (61)\\
&=\sum_{n=1}^{N}\frac{2\rho_n^2}{a_n^2}\sum_{k=0}^{r_n-1}[a_{n,k}]^2\,\mathbb{E}\left[\left\|g_n(\boldsymbol{\omega}_n^{(t,k)})-\nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\right\|^2\right] && (62)\\
&\le 2\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|^2}{\|\boldsymbol{a}_n\|_1^2}, && (63)
\end{aligned}
$$

where equation (62) is derived using Lemma 6 and equation (63) is derived from Assumption 3.

Until now, we can bound T_1 as

$$
T_1\le 2\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|^2}{\|\boldsymbol{a}_n\|_1^2}+2\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]. \tag{64}
$$

2) Bounding T_2 in equation (55)*:* Applying distributive property on T_2 in equation (55) , we can rewrite T_2 as

$$
\begin{aligned}
T_2&=\mathbb{E}\left[\left\langle\nabla F(\boldsymbol{\omega}^{(t,0)}),\sum_{n=1}^{N}\rho_n(\boldsymbol{d}_n^{(t)}-\boldsymbol{h}_n^{(t)})\right\rangle\right]\\
&\quad+\mathbb{E}\left[\left\langle\nabla F(\boldsymbol{\omega}^{(t,0)}),\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\rangle\right] && (65)\\
&=\mathbb{E}\left[\left\langle\nabla F(\boldsymbol{\omega}^{(t,0)}),\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\rangle\right] && (66)\\
&=\frac{1}{2}\left\|\nabla F(\boldsymbol{\omega}^{(t)})\right\|^2+\frac{1}{2}\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]\\
&\quad-\frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})-\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]. && (67)
\end{aligned}
$$

where equation (66) is derived from equation (51) and equation (67) follows the fact $2\langle a,b\rangle=\|a\|^2+\|b\|^2-\|a-b\|^2$.

3) Intermediate result: Substituting the bound for T_1 in equation (64) and T_2 in equation (67) back into (55), we have

$$
\begin{aligned}
&\mathbb{E}[F(\boldsymbol{\omega}^{(t+1,0)})]-F(\boldsymbol{\omega}^{(t,0)})\le\frac{(\sum_{n=1}^{N}\rho_n r_n)^2\eta^2L}{2}\cdot\\
&(2\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|^2}{\|\boldsymbol{a}_n\|_1^2}+2\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right])\\
&-(\sum_{n=1}^{N}\rho_n r_n)\eta(\frac{1}{2}\left\|\nabla F(\boldsymbol{\omega}^{(t)})\right\|^2+\frac{1}{2}\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]\\
&-\frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})-\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]). && (68)
\end{aligned}
$$

To eliminate the term $\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right])$ in equation (68) , we assume that

$$
(\sum_{n=1}^{N}\rho_n r_n)\eta L\le 1/2. \tag{69}
$$

After applying equation (69) and dividing by $\eta(\sum_{n=1}^{N}\rho_n r_n)$, equation (68) can be written as

$$
\begin{aligned}
&\frac{\mathbb{E}\left[F(\boldsymbol{\omega}^{(t+1,0)})\right]-F(\boldsymbol{\omega}^{(t,0)})}{\eta(\sum_{n=1}^{N}\rho_n r_n)}\\
&\le(\sum_{n=1}^{N}\rho_n r_n)\eta L(\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|^2}{\|\boldsymbol{a}_n\|_1^2}+\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right])\\
&-(\frac{1}{2}\left\|\nabla F(\boldsymbol{\omega}^{(t)})\right\|^2+\frac{1}{2}\mathbb{E}\left[\left\|\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]\\
&-\frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})-\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]) && (70)\\
&\le(\sum_{n=1}^{N}\rho_n r_n)\eta L\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|_2^2}{\|\boldsymbol{a}_n\|_1^2}-\frac{1}{2}\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2\\
&+\frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})-\sum_{n=1}^{N}\rho_n\boldsymbol{h}_n^{(t)}\right\|^2\right]. && (71)
\end{aligned}
$$

Further, after applying equation (2) and Jensen's inequality $\left\|\sum_{n=1}^{N}\rho_n z_n\right\|^2\le\sum_{n=1}^{N}\rho_n\|z_n\|^2$ on equation (71), we have

$$
\begin{aligned}
&\frac{\mathbb{E}\left[F(\boldsymbol{\omega}^{(t+1,0)})\right]-F(\boldsymbol{\omega}^{(t,0)})}{\eta(\sum_{n=1}^{N}\rho_n r_n)}\\
&\le(\sum_{n=1}^{N}\rho_n r_n)\eta L\sigma^2\sum_{n=1}^{N}\frac{\rho_n^2\|\boldsymbol{a}_n\|_2^2}{\|\boldsymbol{a}_n\|_1^2}-\frac{1}{2}\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2\\
&+\frac{1}{2}\underbrace{\sum_{n=1}^{N}\rho_n\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})-\boldsymbol{h}_n^{(t)}\right\|^2\right]}_{T_3}. && (72)
\end{aligned}
$$

Before we delve into equation (72), we propose the following assumption to bound the accumulation vector.

Assumption 6. *All elements in the accumulation vector $\boldsymbol{a}_n(k)$, in which $k \in [1, r_n]$, $\forall k$, are upper bounded by Λ, that is*

$$\Lambda = \max_{n,s,k} a_{n,s}(k). \tag{73}$$

Also, $||\boldsymbol{a}_n(k)||_p \leq ||\boldsymbol{a}_n(k+1)||$ for $p = \{1, 2\}$.

In order to bound term T_3 in (72), we can use the following lemma 7 [14].

Lemma 7. *The difference between the locally averaged gradient and the server gradient $\nabla F_n(\boldsymbol{\omega}^{(t,0)})$ can be bounded as follows:*

$$\begin{aligned}&\sum_{n=1}^{N} \rho_n \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\|^2] \\ &\leq \frac{D[\sum_{n=1}^{N} \rho_n (1+\beta_n)^2]}{1-D} \mathbb{E}[\|\nabla F(\boldsymbol{\omega}^{(t,0)})\|^2] + \frac{2\eta^2 L^2 \sigma^2 B}{1-D},\end{aligned} \tag{74}$$

where $B = \Lambda \sum_{n=1}^{N} \frac{\rho_n (r_n - 1) \|\boldsymbol{a}_n\|_2^2}{a_n}$. The proof of Lemma 7 is given in Appendix B below in this online appendix.

4) Final results: Substituting (74) back into (72), we have

$$\begin{aligned}&\frac{\mathbb{E}\left[F(\boldsymbol{\omega}^{(t+1,0)})\right] - F(\boldsymbol{\omega}^{(t,0)})}{\eta(\sum_{n=1}^{N} \rho_n r_n)} \\ &\leq (\frac{D[\sum_{n=1}^{N} \rho_n (1+\beta_n)^2]}{2(1-D)} - \frac{1}{2}) \left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2 \\ &+ (\sum_{n=1}^{N} \rho_n r_n) \eta L \sigma^2 \sum_{n=1}^{N} \frac{\rho_n^2 \|\boldsymbol{a}_n\|_2^2}{\|\boldsymbol{a}_n\|_1^2} + \frac{\eta^2 L^2 \sigma^2 B}{1-D}.\end{aligned} \tag{75}$$

Suppose $1 - D(1 + \sum_{n=1}^{N} \rho_n (1+\beta_n)^2) > 0$, then Equation (75) is equivalent to

$$\begin{aligned}&\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2 \leq \frac{2(1-D)}{1 - D(1 + \sum_{n=1}^{N} \rho_n (1+\beta_n)^2)} \cdot \\ &(\frac{F(\boldsymbol{\omega}^{(t,0)}) - \mathbb{E}\left[F(\boldsymbol{\omega}^{(t+1,0)})\right]}{\eta(\sum_{n=1}^{N} \rho_n r_n)} + \frac{\eta^2 L^2 \sigma^2 B}{1-D} \\ &+ (\sum_{n=1}^{N} \rho_n r_n) \eta L \sigma^2 \sum_{n=1}^{N} \frac{\rho_n^2 \|\boldsymbol{a}_n\|_2^2}{\|\boldsymbol{a}_n\|_1^2}).\end{aligned} \tag{76}$$

Taking the average across all rounds, we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2\right] \leq \frac{2(1-D)}{1 - D(1 + \sum_{n=1}^{N} \rho_n (1+\beta_n)^2)} \cdot \tag{77}$$

$$(\frac{F(\boldsymbol{\omega}^{(T,0)}) - F(\boldsymbol{\omega}^{(0,0)})}{\eta(\sum_{n=1}^{N} \rho_n r_n) T} + \eta L \sigma^2 A + \frac{\eta^2 L^2 \sigma^2 B}{1-D}), \tag{78}$$

where $A = (\sum_{n=1}^{N} \rho_n r_n) \sum_{n=1}^{N} \frac{\sigma_n^2 \rho_n^2 \|\boldsymbol{a}_n\|_2^2}{\|\boldsymbol{a}_n\|_1^2}$.

Since $\min \left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2$, we have

$$\min_{t \in [T]} \mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2\right] \leq \frac{2(1-D)}{1 - D(1 + \sum_{n=1}^{N} \rho_n (1+\beta_n)^2)} \cdot \tag{79}$$

$$(\frac{F(\boldsymbol{\omega}^{(T,0)}) - F(\boldsymbol{\omega}^{(0,0)})}{\eta(\sum_{n=1}^{N} \rho_n r_n) T} + \eta L \sigma^2 A + \frac{\eta^2 L^2 \sigma^2 B}{1-D}). \tag{80}$$

By setting $\eta = \frac{N}{\sqrt{T \sum_{n=1}^{N} r_n}}$, $\min_{t \in [T]} \mathbb{E}\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2$ in equation (80) will be upper bounded by

$$\begin{aligned}&\min_{t \in [T]} \mathbb{E}\left[\left\|\nabla F(\boldsymbol{\omega}^{(t,0)})\right\|^2\right] \leq \frac{2(1-D)}{1 - D(1 + \sum_{n=1}^{N} \rho_n (1+\beta_n)^2)} \cdot \\ &(\frac{[F(\boldsymbol{\omega}^{(T,0)}) - F(\boldsymbol{\omega}^{(0,0)})]\sqrt{\sum_{n=1}^{N} r_n}}{N(\sum_{n=1}^{N} \rho_n r_n)\sqrt{T}} + \frac{N L \sigma^2 A}{\sqrt{T \sum_{n=1}^{N} r_n}} \\ &+ \frac{N^2 L^2 \sigma^2 B}{(1-D) T \sum_{n=1}^{N} r_n}).\end{aligned} \tag{81}$$

Thus, we conclude that the algorithm converges to a stationary point of $F(\boldsymbol{\omega})$ in a rate of

$$\mathcal{O}(\frac{1}{(G - \sum_{n=1}^{N} \rho_n (1+\beta_n)^2)\sqrt{T \sum_{n=1}^{N} r_n}}). \tag{82}$$

where G is a constant that makes $G > \sum_{n=1}^{N} \rho_n (1+\beta_n)^2$. Here, we complete the proof of Theorem 1. □

B. Proof of Lemma 7

Proof. Substituting the definition of $\boldsymbol{h}_n^{(t)}$ in equation (50) into $\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\right\|^2\right]$ in T_3, we can derive that

$$\mathbb{E}\left[\left\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\right\|^2\right]$$

$$= \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \frac{1}{a_n} \sum_{k=0}^{r_n - 1} a_{n,k} \nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\|^2] \tag{83}$$

$$= \mathbb{E}[\|\frac{1}{a_n} \sum_{k=0}^{r_n - 1} a_{n,k} \big(\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\big)\|^2] \tag{84}$$

$$\leq \frac{1}{a_n} \sum_{k=0}^{r_n - 1} \{a_{n,k} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\|^2]\} \tag{85}$$

$$\leq \frac{L^2}{a_n} \sum_{k=0}^{r_n - 1} \{a_{n,k} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2]\} \tag{86}$$

$$\leq \frac{L^2 \Lambda}{a_n} \sum_{k=0}^{r_n - 1} \{\mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2]\}, \tag{87}$$

where equation (84) is generated from $a_n = \|\boldsymbol{a}_n\|_1$, equation (85) uses Jensen's inequality, equation (86) follows Assumption 1 and equation (87) is derived from equation (73).

Now, we turn to bound the right-hand side in equation (87). Plugging into the local update rule in equation (49), and using the fact $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] \\
&= \mathbb{E}[\| - \eta \sum_{s=0}^{k-1} a_{n,s}(k) g_n(\boldsymbol{\omega}^{(t,s)})\|^2] && (88)\\
&= \eta^2 \mathbb{E}[\| \sum_{s=0}^{k-1} a_{n,s}(k) g_n(\boldsymbol{\omega}^{(t,s)})\|^2] && (89)\\
&\le 2\eta^2 \mathbb{E}[\| \sum_{s=0}^{k-1} a_{n,s}(k)\big(g_n(\boldsymbol{\omega}_n^{(t,s)}) - \nabla F_n(\boldsymbol{\omega}_n^{(t,s)})\big)\|^2] \\
&+ 2\eta^2 \mathbb{E}[\| \sum_{s=0}^{k-1} a_{n,s}(k) \nabla F_n(\boldsymbol{\omega}_n^{(t,s)})\|^2]. && (90)
\end{aligned}$$

Applying Assumption 3 and Lemma 6 to the first term in equation (90), we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] \\
&\le 2\eta^2 \sigma^2 \sum_{s=0}^{k-1} [a_{n,s}(k)]^2 \\
&+ 2\eta^2 \mathbb{E}[\| \sum_{s=0}^{k-1} a_{n,s}(k) \nabla F_n(\boldsymbol{\omega}_n^{(t,s)})\|^2] && (91)\\
&\le 2\eta^2 \sigma^2 \sum_{s=0}^{k-1} [a_{n,s}(k)]^2 \\
&+ 2\eta^2 [\sum_{s=0}^{k-1} a_{n,s}(k)] \sum_{s=0}^{k-1} a_{n,s}(k) \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}_n^{(t,s)})\|^2] && (92)\\
&\le 2\eta^2 \sigma^2 \sum_{s=0}^{k-1} [a_{n,s}(k)]^2 \\
&+ 2\eta^2 \Lambda [\sum_{s=0}^{k-1} a_{n,s}(k)] \sum_{s=0}^{r_n-1} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}_n^{(t,s)})\|^2], && (93)
\end{aligned}$$

where (92) follows from Jensen's inequality, and (93) uses Assumption 6.

From Assumption 6 where $\|\boldsymbol{a}_n(k)\| \le \|\boldsymbol{a}_n(r_n)\| = \|\boldsymbol{a}_n\|$, we have

$$\sum_{k=0}^{r_n-1} \left[\sum_{s=0}^{k-1} [a_{n,s}(k)]^2 \right] = \sum_{k=0}^{r_n-1} \|\boldsymbol{a}_n(k)\|_2^2 \le (r_n - 1)\|\boldsymbol{a}_n\|_2^2, \tag{94}$$

and

$$\sum_{k=0}^{r_n-1} \left[\sum_{s=0}^{k-1} [a_{n,s}(k)] \right] = \sum_{k=0}^{r_n-1} \|\boldsymbol{a}_n(k)\|_1 \le (r_n - 1)\|\boldsymbol{a}_n\|_1. \tag{95}$$

After substituting the result in equation (93), (94) and (95) into the term $\frac{L^2\Lambda}{a_n} \sum_{k=0}^{r_n-1} \{\mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2]\}$ in equation (87), we get that

$$\begin{aligned}
&\frac{L^2\Lambda}{a_n} \sum_{k=0}^{r_n-1} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] \\
&\le \frac{2\eta^2 L^2 \Lambda \sigma^2 (r_n - 1)\|\boldsymbol{a}_n\|_2^2}{a_n} \\
&+ 2\eta^2 L^2 \Lambda^2 (r_n - 1) \sum_{k=0}^{r_n-1} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\|^2] && (96)
\end{aligned}$$

Applying the fact $\|a+b\|^2 \le 2\,\|a\|^2 + 2\,\|b\|^2$ and Assumption 1, the second term in equation (96) can be bounded by

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}_n^{(t,k)})\|^2] \\
&\le 2\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}_n^{(t,k)}) - \nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2] \\
&+ 2\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2] && (97)\\
&\le 2L^2 \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] + 2\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2]. && (98)
\end{aligned}$$

Substituting equation (98) into the second term in (96), we get that

$$\begin{aligned}
&\frac{L^2\Lambda}{a_n} \sum_{k=0}^{r_n-1} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] \\
&\le \frac{2\eta^2 L^2 \Lambda \sigma^2 (r_n - 1)\|\boldsymbol{a}_n\|_2^2}{a_n} \\
&+ 4\eta^2 L^4 \Lambda^2 (r_n - 1) \sum_{k=0}^{r_n-1} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] \\
&+ 4\eta^2 L^2 \Lambda^2 r_n (r_n - 1) \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2]. && (99)
\end{aligned}$$

After minor rearranging, it follows that

$$\begin{aligned}
&\frac{L^2\Lambda}{a_n} \sum_{k=0}^{r_n-1} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] && (100)\\
&\le \frac{2\eta^2 L^2 \Lambda \sigma^2}{1 - 4\eta^2 L^2 \Lambda (r_n - 1) a_n} \frac{(r_n - 1)\|\boldsymbol{a}_n\|_2^2}{a_n} \\
&+ \frac{4\eta^2 L^2 \Lambda^2 r_n (r_n - 1)}{1 - 4\eta^2 L^2 \Lambda (r_n - 1) a_n} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2]. && (101)
\end{aligned}$$

As given in equation (73), we have $a_n \le \Lambda r_n$. Thus, the upper bound of $\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\|^2]$ in equation (87) can be written as

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\|^2] \\
&\le \frac{L^2\Lambda}{a_n} \sum_{k=0}^{r_n-1} \mathbb{E}[\|\boldsymbol{\omega}^{(t,0)} - \boldsymbol{\omega}_n^{(t,k)}\|^2] && (102)\\
&\le \frac{2\eta^2 L^2 \Lambda \sigma^2}{1 - 4\eta^2 L^2 \Lambda^2 r_n (r_n - 1)} \frac{(r_n - 1)\|\boldsymbol{a}_n\|_2^2}{a_n} \\
&+ \frac{4\eta^2 L^2 \Lambda^2 r_n (r_n - 1)}{1 - 4\eta^2 L^2 \Lambda^2 r_n (r_n - 1)} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2]. && (103)
\end{aligned}$$

Define

$$D \triangleq 4\eta^2 L^2 \Lambda^2 \max_n r_n (r_n - 1) < 1, \tag{104}$$

we can simplify (103) as follows

$$\begin{aligned}
&\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)}) - \boldsymbol{h}_n^{(t)}\|^2] \le \frac{D}{1-D} \mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2] \\
&+ \frac{2\eta^2 L^2 \Lambda \sigma^2}{1 - D} \frac{(r_n - 1)\|\boldsymbol{a}_n\|_2^2}{a_n}. && (105)
\end{aligned}$$

Then, taking the weighted average across all clients and applying Assumption 4, equation (105) can be rewritten as

$$
\begin{aligned}
&\sum_{n=1}^{N}\rho_n\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})-\boldsymbol{h}_n^{(t)}\|^2] \\
&\leq \frac{D}{1-D}\sum_{n=1}^{N}\rho_n\mathbb{E}[\|\nabla F_n(\boldsymbol{\omega}^{(t,0)})\|^2] \\
&+\frac{2\eta^2L^2\Lambda\sigma^2}{1-D}\sum_{n=1}^{N}\frac{\rho_n(r_n-1)\|\boldsymbol{a}_n\|_2^2}{a_n} \qquad (106)\\
&\leq \frac{D[\sum_{n=1}^{N}\rho_n(1+\beta_n)^2]}{1-D}\mathbb{E}[\|\nabla F(\boldsymbol{\omega}^{(t,0)})\|^2]+\frac{2\eta^2L^2\sigma^2B}{1-D}, \qquad (107)
\end{aligned}
$$

where $B \triangleq \Lambda\sum_{n=1}^{N}\frac{\rho_n(r_n-1)\|\boldsymbol{a}_n\|_2^2}{a_n}$. $\square$