

Proof:

Prior to delving into the convergence analysis, we first introduce widely accepted assumptions, as outlined below:

- **Assumption 1:** $\nabla F(\mathbf{w})$ is uniformly L -Lipschitz continuous in reference to \mathbf{w} , which is represented as

$$\|\nabla F(\mathbf{w}^{n+1}) - \nabla F(\mathbf{w}^n)\| \leq L \|\mathbf{w}^{n+1} - \mathbf{w}^n\|, \quad (\text{A.1})$$

where L is the Lipschitz constant associated with $F(\cdot)$.

- **Assumption 2:** $F(\mathbf{w})$ is γ -strongly convex, satisfying

$$F(\mathbf{w}^{n+1}) \geq F(\mathbf{w}^n) + (\mathbf{w}^{n+1} - \mathbf{w}^n)^\top \nabla F(\mathbf{w}^n) + \frac{\gamma}{2} \|\mathbf{w}^{n+1} - \mathbf{w}^n\|^2, \quad (\text{A.2})$$

where γ is determined by $F(\cdot)$.

- **Assumption 3:** $\nabla F(\mathbf{w})$ is twice-continuously differentiable. Given Assumptions 1 and 2, we can obtain

$$\gamma \mathbf{I} \preceq \nabla^2 F(\mathbf{w}) \preceq L \mathbf{I}, \quad (\text{A.3})$$

where \mathbf{I} denotes an identity matrix.

- **Assumption 4:** The second moments of local gradient and parameters are constrained by

$$\mathbb{E} \left\{ \|\nabla f(\mathbf{w})\|^2 \right\} \leq A^2, \quad (\text{A.4})$$

and

$$\mathbb{E} \left\{ \|\mathbf{w}\|^2 \right\} \leq D^2. \quad (\text{A.5})$$

- **Assumption 5:** The stochastic gradients are unbiased, which can be represented as

$$\mathbb{E}\{g(\mathbf{w})\} = \nabla F(\mathbf{w}). \quad (\text{A.6})$$

It should be noted that the most of loss functions readily meet these assumptions [1], [2].

For simplicity, we use $\hat{\mathbf{g}}^n(\hat{\mathbf{w}})$ to represent $\hat{\mathbf{g}}^n(\hat{\mathbf{w}}; r_z^n)$, and $\bar{\mathbf{g}}^n(\hat{\mathbf{w}})$ represents $\bar{\mathbf{g}}^n(\hat{\mathbf{w}}; r_z^n, b_z^n)$.

To facilitate the following analysis, we introduce two auxiliary variables as

$$\boldsymbol{\lambda}_1^n = \nabla F(\mathbf{w}^n) - \bar{\mathbf{g}}^n(\hat{\mathbf{w}}) \quad (\text{A.7})$$

and

$$\boldsymbol{\lambda}_2^n = \bar{\mathbf{g}}^n(\hat{\mathbf{w}}) - \hat{\mathbf{g}}^n(\hat{\mathbf{w}}), \quad (\text{A.8})$$

respectively. Hence, Eq. (16) can be rewritten as

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \eta (\nabla F(\mathbf{w}^n) - \boldsymbol{\lambda}_1^n). \quad (\text{A.9})$$

Furthermore, we rewrite $F(\mathbf{w}^{n+1})$ as the expression of its second-order Taylor expansion, which can be represented as

$$\begin{aligned} F(\mathbf{w}^{n+1}) &\leq F(\mathbf{w}^n) + (\nabla F(\mathbf{w}^n))^\top (\mathbf{w}^{n+1} - \mathbf{w}^n) \\ &\quad + \frac{\nabla^2 F(\mathbf{w}^n)}{2} \|\mathbf{w}^{n+1} - \mathbf{w}^n\|^2 \\ &\stackrel{(a)}{\leq} F(\mathbf{w}^n) + (\nabla F(\mathbf{w}^n))^\top (\mathbf{w}^{n+1} - \mathbf{w}^n) \\ &\quad + \frac{L}{2} \|\mathbf{w}^{n+1} - \mathbf{w}^n\|^2 \\ &\stackrel{(b)}{\leq} F(\mathbf{w}^n) - \eta (\nabla F(\mathbf{w}^n))^\top (\nabla F(\mathbf{w}^n) - \boldsymbol{\lambda}_1^n) \\ &\quad + \frac{L\eta^2}{2} \|\nabla F(\mathbf{w}^n) - \boldsymbol{\lambda}_1^n\|^2 \\ &\leq F(\mathbf{w}^n) - \eta \|\nabla F(\mathbf{w}^n)\|^2 + \eta (\boldsymbol{\lambda}_1^n)^\top \nabla F(\mathbf{w}^n) \\ &\quad + \frac{L\eta^2}{2} \|\nabla F(\mathbf{w}^n) - \boldsymbol{\lambda}_1^n\|^2, \end{aligned} \quad (\text{A.10})$$

where inequality (a) stems from Eq. (A.3), and inequality (b) is due to Eq. (A.9). Given learning rate $\eta = \frac{1}{L}$, we have

$$\begin{aligned} &\mathbb{E} \{ F(\mathbf{w}^{n+1}) \} \\ &\leq \mathbb{E} \left\{ F(\mathbf{w}^n) - \frac{1}{L} \|\nabla F(\mathbf{w}^n)\|^2 + \frac{1}{2L} \|\nabla F(\mathbf{w}^n)\|^2 \right. \\ &\quad \left. + \frac{1}{2L} \|\boldsymbol{\lambda}_1^n\|^2 + \frac{1}{L} (\boldsymbol{\lambda}_1^n)^\top \nabla F(\mathbf{w}^n) - \frac{1}{L} (\boldsymbol{\lambda}_1^n)^\top \nabla F(\mathbf{w}^n) \right\} \\ &\leq \mathbb{E} \left\{ F(\mathbf{w}^n) - \frac{1}{2L} \|\nabla F(\mathbf{w}^n)\|^2 + \frac{1}{2L} \|\boldsymbol{\lambda}_1^n\|^2 \right\} \\ &\stackrel{(c)}{\leq} \mathbb{E} \{ F(\mathbf{w}^n) \} - \frac{\gamma}{2L} + \frac{1}{2L} \mathbb{E} \left\{ \|\boldsymbol{\lambda}_1^n\|^2 \right\}, \end{aligned} \quad (\text{A.11})$$

where (c) is obtained by Eq. (A.3). Due to Eq. (A.7) and Eq. (A.8), we have

$$\begin{aligned} \mathbb{E} \left\{ \|\boldsymbol{\lambda}_1^n\|^2 \right\} &= \mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n) - \bar{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2 \right\} \\ &= \mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n) - \hat{\mathbf{g}}^n(\hat{\mathbf{w}}) - \boldsymbol{\lambda}_2^n\|^2 \right\} \\ &\leq 2\mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n) - \hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2 \right\} + 2\mathbb{E} \left\{ \|\boldsymbol{\lambda}_2^n\|^2 \right\} \\ &\leq 2\mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n)\|^2 \right\} + 2\mathbb{E} \left\{ \|\hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2 \right\} \\ &\quad - 4\mathbb{E} \left\{ (\nabla F(\mathbf{w}^n))^\top \hat{\mathbf{g}}^n(\hat{\mathbf{w}}) \right\} + 2\mathbb{E} \left\{ \|\boldsymbol{\lambda}_2^n\|^2 \right\}. \end{aligned} \quad (\text{A.12})$$

Because of $\mathbb{E} \{ \hat{\mathbf{g}}^n(\hat{\mathbf{w}}) \} = \nabla F(\hat{\mathbf{w}}^n)$ referring to Assumption 5 and $\mathbb{E} \left\{ (\nabla F(\mathbf{w}^n))^\top \hat{\mathbf{g}}^n(\hat{\mathbf{w}}) \right\} = \mathbb{E} \left\{ \nabla F(\mathbf{w}^n) \right\}^\top \mathbb{E} \{ \hat{\mathbf{g}}^n(\hat{\mathbf{w}}) \} + \text{Tr}(\text{Cov}((\nabla F(\mathbf{w}^n))^\top, \hat{\mathbf{g}}^n(\hat{\mathbf{w}})))$, we can obtain

$$\begin{aligned} \mathbb{E} \left\{ \|\boldsymbol{\lambda}_1^n\|^2 \right\} &\leq 2\mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n)\|^2 \right\} + 2\mathbb{E} \left\{ \|\hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2 \right\} \\ &\quad + 2\mathbb{E} \left\{ \|\boldsymbol{\lambda}_2^n\|^2 \right\} - 4\mathbb{E} \left\{ (\nabla F(\mathbf{w}^n))^\top \nabla F(\hat{\mathbf{w}}^n) \right\} \\ &\leq 2\mathbb{E} \left\{ \|\nabla F(\mathbf{w}^n) - \nabla F(\hat{\mathbf{w}}^n)\|^2 \right\} + 2\mathbb{E} \left\{ \|\boldsymbol{\lambda}_2^n\|^2 \right\} \\ &\quad + 2\mathbb{E} \left\{ \|\hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2 \right\} - 2\mathbb{E} \left\{ \|\nabla F(\hat{\mathbf{w}}^n)\|^2 \right\}. \end{aligned} \quad (\text{A.13})$$

In the following, we investigate the upper bounds of $\mathbb{E}\{\|\hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2\}$, $\mathbb{E}\{\|\boldsymbol{\lambda}_2^n\|^2\}$, and $\mathbb{E}\{\|\nabla F(\mathbf{w}^n) - \nabla F(\hat{\mathbf{w}}^n)\|^2\}$, respectively. Firstly,

$$\begin{aligned} \mathbb{E}\{\|\hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2\} &\triangleq E\left\{\left\|\frac{\sum_{z=1}^Z N_z \hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)}{\sum_{z=1}^Z N_z}\right\|^2\right\} \\ &\stackrel{(d)}{\leq} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^Z \|N_z\|^2\right) \left(\sum_{z=1}^Z \|\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)\|^2\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2}\right\} \quad (\text{A.14}) \\ &\stackrel{(e)}{\leq} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^Z \|N_z\|^2\right) \left(\sum_{z=1}^Z \|\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)\|^2\right)}{\sum_{z=1}^Z \|N_z\|^2}\right\} \\ &= \mathbb{E}\left\{\sum_{z=1}^Z \|\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)\|^2\right\} \stackrel{(f)}{\leq} UA^2, \end{aligned}$$

where inequality (d) arises from Cauchy-Buniakowsky-Schwarz inequality (i.e., $\sum_{i=1}^n \|a_i\|^2 \sum_{i=1}^n \|b_i\|^2 \geq \sum_{i=1}^n \|a_i b_i\|^2$), while inequality (e) follows from the fact that $\sum_{i=1}^n a_i^2 \leq (\sum_{i=1}^n a_i)^2$, and inequality (f) is derived from Assumption 4. Let U_1 represent the set of the devices without transmission failure, while U_2 represent the set of devices with transmission failure. Furthermore, the upper bound of $\mathbb{E}\{\|\boldsymbol{\lambda}_2^n\|^2\}$ can be represented as

$$\begin{aligned} \mathbb{E}\{\|\boldsymbol{\lambda}_2^n\|^2\} &\triangleq \mathbb{E}\left[\|\bar{\mathbf{g}}^n(\hat{\mathbf{w}}) - \hat{\mathbf{g}}^n(\hat{\mathbf{w}})\|^2\right] \\ &= \mathbb{E}\left\{\left\|\frac{\sum_{z=1}^Z N_z (Q(\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)) - \hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n))}{\sum_{z=1}^Z N_z}\right\|^2\right\} \\ &\stackrel{(g)}{\leq} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^Z \|N_z\|^2\right) \left(\sum_{z=1}^Z \|Q(\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)) - \hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)\|^2\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2}\right\} \\ &\stackrel{(i)}{\leq} \frac{\left(\sum_{z=1}^Z \|N_z\|^2\right) \left(\sum_{z=1}^Z \mathbb{E}\left\{\|Q(\hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)) - \hat{\mathbf{g}}_z^n(\hat{\mathbf{w}}_z^n)\|^2\right\}\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2} L^2 \\ &\stackrel{(i)}{\leq} \frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \sum_{z=1}^Z \sum_{v=1}^V \frac{\left(\bar{g}_{z,v}^n - \underline{g}_{z,v}^n\right)^2}{4(2b_z^n - 1)^2}}{\left\|\sum_{z=1}^Z N_z\right\|^2} L^2 \triangleq \Lambda_1^n, \quad (\text{A.15}) \end{aligned}$$

where inequality (g) is due to Cauchy-Buniakowsky-Schwarz inequality. For convenience, we use $\mathbb{E}\{\Delta\}$ to represent $\mathbb{E}\{\|\nabla F(\mathbf{w}^n) - \nabla F(\hat{\mathbf{w}}^n)\|^2\}$, and the upper bound of

$\mathbb{E}\{\Delta\}$ can be obtained by

$$\begin{aligned} \mathbb{E}\{\Delta\} &= \mathbb{E}\left\{\left\|\frac{\sum_{z=1}^Z N_z (\nabla F_z(\mathbf{w}_z^n) - \nabla F_z(\hat{\mathbf{w}}_z^n))}{\sum_{z=1}^Z N_z}\right\|^2\right\} L^2 \\ &\stackrel{(k)}{\leq} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \left(\sum_{z=1}^Z \|\nabla F_z(\mathbf{w}_z^n) - \nabla F_z(\hat{\mathbf{w}}_z^n)\|^2\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2}\right\} L^2 \\ &\stackrel{(j)}{\leq} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \left(\sum_{z=1}^Z \|\mathbf{w}_z^n - \hat{\mathbf{w}}_z^n\|^2\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2}\right\} L^2 \\ &\stackrel{(j)}{\leq} \frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \left(\sum_{z=1}^Z \mathbb{E}\{\|\mathbf{w}_z^n - \hat{\mathbf{w}}_z^n\|^2\}\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2} L^2 \\ &\stackrel{(j)}{\leq} \frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \left(\sum_{z=1}^Z r_z^n D^2\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2} L^2 \\ &= \frac{\left(\sum_{z=1}^U \|N_z\|^2\right) \left(\sum_{z=1}^Z r_z^n\right)}{\left\|\sum_{z=1}^Z N_z\right\|^2} L^2 D^2 \triangleq L^2 D^2 \Lambda_2^n, \quad (\text{A.16}) \end{aligned}$$

where inequality (j) is from the Assumption 1, while equality (k) is because of Cauchy-Buniakowsky-Schwarz inequality.

Therefore, substituting Eq. (A.14), Eq. (A.15), and Eq. (A.16) into Eq. (A.13), we can obtain

$$\begin{aligned} \mathbb{E}\{\|\boldsymbol{\lambda}_1^n\|^2\} &\leq 2L^2 D^2 \Lambda_2^n + 2UA^2 \\ &\quad - 2\mathbb{E}\{\|\nabla F(\hat{\mathbf{w}}^n)\|^2\} + 2\Lambda_1^n. \quad (\text{A.17}) \end{aligned}$$

Furthermore, let we substitute Eq. (A.17) into Eq. (A.11), we have

$$\begin{aligned} \mathbb{E}\{F(\mathbf{w}^{n+1})\} &\leq \mathbb{E}\{F(\mathbf{w}^n)\} - \frac{\gamma}{2L} - \frac{1}{L} \mathbb{E}\{\|\nabla F(\hat{\mathbf{w}}^n)\|^2\} \\ &\quad + \frac{ZA^2}{L} + LD^2 \Lambda_2^n + \frac{\Lambda_1^n}{L}. \quad (\text{A.18}) \end{aligned}$$

Rearranging Eq. (A.18), we can obtain

$$\begin{aligned} \mathbb{E}\{\|\nabla F(\hat{\mathbf{w}}^n)\|^2\} &\leq L \mathbb{E}\{F(\mathbf{w}^n) - F(\mathbf{w}^{n+1})\} + UA^2 - \frac{\gamma}{2} \\ &\quad + L^2 D^2 \Lambda_2^n + \Lambda_1^n. \quad (\text{A.19}) \end{aligned}$$

Summing up the above terms from $n = 0$ to Ω and dividing both sides by the total number of iterations, we can obtain

$$\begin{aligned} \frac{1}{\Omega+1} \sum_{n=0}^{\Omega} \mathbb{E}\{\|\nabla F(\hat{\mathbf{w}}^n)\|^2\} &\leq \frac{L}{\Omega+1} \mathbb{E}\{F(\mathbf{w}^0) - F(\mathbf{w}^*)\} \\ &\quad - \frac{\gamma}{2} + UA^2 + \frac{L^2 D^2}{\Omega+1} \sum_{n=0}^{\Omega} \Lambda_2^n + \frac{1}{\Omega+1} \sum_{n=0}^{\Omega} \Lambda_1^n. \quad (\text{A.20}) \end{aligned}$$

Thus, we obtain the average ℓ_2 -norm of the gradients as

$$\begin{aligned}
\frac{1}{\Omega+1} \mathbb{E} \left\{ \|\nabla F(\hat{\mathbf{w}}^n)\|^2 \right\} &\leq \frac{1}{\Omega+1} \sum_{n=0}^{\Omega} \mathbb{E} \left\{ \|\nabla F(\hat{\mathbf{w}}^n)\|^2 \right\} \\
&\leq \frac{L}{\Omega+1} \mathbb{E} \{ F(\mathbf{w}^0) - F(\mathbf{w}^*) \} \\
&\quad - \frac{\Lambda}{2} + UA^2 + \frac{L^2 D^2}{\Omega+1} \sum_{n=0}^{\Omega} \Lambda_2^n \\
&\quad + \frac{1}{\Omega+1} \sum_{n=0}^{\Omega} \Lambda_1^n,
\end{aligned} \tag{A.21}$$

where \mathbf{w}^* is the optimal model. Let $\Lambda^n = L^2 D^2 \Lambda_2^n + \Lambda_1^n$, and then Eq. (A.21) can be rewritten as

$$\begin{aligned}
\frac{1}{\Omega+1} \mathbb{E} \left\{ \|\nabla F(\hat{\mathbf{w}}^n)\|^2 \right\} &\leq \frac{2L}{\Omega+1} \mathbb{E} \{ F(\mathbf{w}^0) - F(\mathbf{w}^*) \} \\
&\quad - \gamma + ZA^2 + \frac{1}{\Omega+1} \sum_{n=0}^{\Omega} \Lambda^n.
\end{aligned} \tag{A.22}$$

where Λ^n is given by

$$\begin{aligned}
\Lambda^n &= \frac{\sum_{z=1}^Z \|N_z\|^2}{\left\| \sum_{z=1}^Z N_z \right\|^2} \cdot \left(L^2 D^2 \sum_{z=1}^Z r_z^n \right. \\
&\quad \left. + \sum_{z=1}^Z \frac{\sum_{v=1}^V (\bar{g}_{z,v}^n - g_{z,v}^n)^2}{4(2^{b_z^n} - 1)^2} \right)
\end{aligned} \tag{A.23}$$

This completes the proof. \blacksquare

REFERENCES

- [1] X. Fan, Y. Wang, Y. Huo, and Z. Tian, “1-bit compressive sensing for efficient federated learning over the air,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 2139–2155, 2023.
- [2] M. Chen, H. V. Poor, W. Saad, and S. Cui, “Convergence time optimization for federated learning over wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2457–2471, 2021.