Proof:

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

- **Assumption 1:** *$\nabla F(\boldsymbol{w})$ is uniformly L-Lipschitz continuous in reference to $\boldsymbol{w}$, which is represented as*
$$\left\|\nabla F\left(\boldsymbol{w}^{n+1}\right)-\nabla F\left(\boldsymbol{w}^{n}\right)\right\| \leq L\left\|\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right\|, \tag{A.1}$$
where L is the Lipschitz constant associated with $F(\cdot)$.
- **Assumption 2:** *$F(\boldsymbol{w})$ is γ-strongly convex, satisfying*
$$\begin{aligned} F\left(\boldsymbol{w}^{n+1}\right) \geq & F\left(\boldsymbol{w}^{n}\right)+\left(\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right)^{\mathrm{T}} \nabla F\left(\boldsymbol{w}^{n}\right) \\ & +\frac{\gamma}{2}\left\|\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right\|^{2}, \end{aligned} \tag{A.2}$$
where γ is determined by $F(\cdot)$.
- **Assumption 3:** *$\nabla F(\boldsymbol{w})$ is twice-continuously differentiable. Given Assumptions 1 and 2 , we can obtain*
$$\gamma \boldsymbol{I} \preceq \nabla^{2} F(\boldsymbol{w}) \preceq L \boldsymbol{I}, \tag{A.3}$$
where $\boldsymbol{I}$ denotes an identity matrix.
- **Assumption 4:** *The second moments of local gradient and parameters are constrained by*
$$\mathbb{E}\left\{\|\nabla f(\boldsymbol{w})\|^{2}\right\} \leq A^{2}, \tag{A.4}$$
and
$$\mathbb{E}\left\{\|\boldsymbol{w}\|^{2}\right\} \leq D^{2}. \tag{A.5}$$
- **Assumption 5:** *The stochastic gradients are unbiased, which can be represented as*
$$\mathbb{E}\{g(\boldsymbol{w})\}=\nabla F(\boldsymbol{w}). \tag{A.6}$$

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

For simplicity, we use $\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})$ to represent $\hat{\boldsymbol{g}}^{n}\left(\hat{\boldsymbol{w}}; r_{z}^{n}\right)$, and $\overline{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})$ represents $\overline{\boldsymbol{g}}^{n}\left(\hat{\boldsymbol{w}}; r_{z}^{n}, b_{z}^{n}\right)$.

To facilitate the following analysis, we introduce two auxiliary variables as
$$\boldsymbol{\lambda}_{1}^{n}=\nabla F\left(\boldsymbol{w}^{n}\right)-\overline{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}}) \tag{A.7}$$
and
$$\boldsymbol{\lambda}_{2}^{n}=\overline{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})-\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}}), \tag{A.8}$$
respectively. Hence, Eq. (16) can be rewritten as
$$\boldsymbol{w}^{n+1}=\boldsymbol{w}^{n}-\eta\left(\nabla F\left(\boldsymbol{w}^{n}\right)-\boldsymbol{\lambda}_{1}^{n}\right). \tag{A.9}$$

Furthermore, we rewrite $F\left(\boldsymbol{w}^{n+1}\right)$ as the expression of its second-order Taylor expansion, which can be represented as
$$\begin{aligned} F\left(\boldsymbol{w}^{n+1}\right) \leq & F\left(\boldsymbol{w}^{n}\right)+\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top}\left(\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right) \\ & +\frac{\nabla^{2} F\left(\boldsymbol{w}^{n}\right)}{2}\left\|\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right\|^{2} \\ \overset{(a)}{\leq} & F\left(\boldsymbol{w}^{n}\right)+\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top}\left(\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right) \\ & +\frac{L}{2}\left\|\boldsymbol{w}^{n+1}-\boldsymbol{w}^{n}\right\|^{2} \\ \overset{(b)}{\leq} & F\left(\boldsymbol{w}^{n}\right)-\eta\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top}\left(\nabla F\left(\boldsymbol{w}^{n}\right)-\boldsymbol{\lambda}_{1}^{n}\right) \\ & +\frac{L \eta^{2}}{2}\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\boldsymbol{\lambda}_{1}^{n}\right\|^{2} \\ \leq & F\left(\boldsymbol{w}^{n}\right)-\eta\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}+\eta\left(\boldsymbol{\lambda}_{1}^{n}\right)^{\top} \nabla F\left(\boldsymbol{w}^{n}\right) \\ & +\frac{L \eta^{2}}{2}\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\boldsymbol{\lambda}_{1}^{n}\right\|^{2}, \end{aligned} \tag{A.10}$$
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
$$\begin{aligned} & \mathbb{E}\left\{F\left(\boldsymbol{w}^{n+1}\right)\right\} \\ \leq & \mathbb{E}\left\{F\left(\boldsymbol{w}^{n}\right)-\frac{1}{L}\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}+\frac{1}{2L}\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}\right. \\ & \left.+\frac{1}{2L}\left\|\boldsymbol{\lambda}_{1}^{n}\right\|^{2}+\frac{1}{L}\left(\boldsymbol{\lambda}_{1}^{n}\right)^{\top} \nabla F\left(\boldsymbol{w}^{n}\right)-\frac{1}{L}\left(\boldsymbol{\lambda}_{1}^{n}\right)^{\top} \nabla F\left(\boldsymbol{w}^{n}\right)\right\} \\ \leq & \mathbb{E}\left\{F\left(\boldsymbol{w}^{n}\right)-\frac{1}{2L}\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}+\frac{1}{2L}\left\|\boldsymbol{\lambda}_{1}^{n}\right\|^{2}\right\} \\ \overset{(c)}{\leq} & \mathbb{E}\left\{F\left(\boldsymbol{w}^{n}\right)\right\}+\frac{1}{2L} \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{1}^{n}\right\|^{2}\right\}, \end{aligned} \tag{A.11}$$
where (c) is obtained by Eq. (A.3). Due to Eq. (A.7) and Eq. (A.8), we have
$$\begin{aligned} \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{1}^{n}\right\|^{2}\right\} & =\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\overline{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\|^{2}\right\} \\ & =\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})-\boldsymbol{\lambda}_{2}^{n}\right\|^{2}\right\} \\ & \leq 2 \mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\|^{2}\right\}+2 \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{2}^{n}\right\|^{2}\right\} \\ & \leq 2 \mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}\right\}+2 \mathbb{E}\left\{\left\|\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\|^{2}\right\} \\ & -4 \mathbb{E}\left\{\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top} \hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\}+2 \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{2}^{n}\right\|^{2}\right\}. \end{aligned} \tag{A.12}$$
Because of $\mathbb{E}\left\{\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\}=\nabla F\left(\hat{\boldsymbol{w}}^{n}\right)$ referring to Assumption 5 and $\mathbb{E}\left\{\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top} \hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\}=E\left\{\nabla F\left(\boldsymbol{w}^{n}\right)\right\}^{\top} \mathbb{E}\left\{\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\}+\operatorname{Tr}(\operatorname{Cov}((\nabla F\left(\boldsymbol{w}^{n}\right))^{\top}, \hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})))$, we can obtain
$$\begin{aligned} \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{1}^{n}\right\|^{2}\right\} & \leq 2 \mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)\right\|^{2}\right\}+2 \mathbb{E}\left\{\left\|\hat{\boldsymbol{g}}^{n}(\hat{\boldsymbol{w}})\right\|^{2}\right\} \\ & +2 \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{2}^{n}\right\|^{2}\right\}-4 \mathbb{E}\left\{\left(\nabla F\left(\boldsymbol{w}^{n}\right)\right)^{\top} \nabla F\left(\hat{\boldsymbol{w}}^{n}\right)\right\} \\ & \leq 2 \mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^{n}\right)-\nabla F\left(\hat{\boldsymbol{w}}^{n}\right)\right\|^{2}\right\}+2 \mathbb{E}\left\{\left\|\boldsymbol{\lambda}_{2}^{n}\right\|^{2}\right\} \\ & +2 \mathbb{E}\left\{\left\|\hat{\boldsymbol{g}^{n}}(\hat{\boldsymbol{w}})\right\|^{2}\right\}-2 \mathbb{E}\left\{\left\|\nabla F\left(\hat{\boldsymbol{w}}^{n}\right)\right\|^{2}\right\}. \end{aligned} \tag{A.13}$$

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

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

where inequality (d) arises from Cauchy-Buniakowsky-Schwarz inequality (i.e., $\sum_{i=1}^{n}\|a_i\|^2\sum_{i=1}^{n}\|b_i\|^2 \ge \sum_{i=1}^{n}\|a_i b_i\|^2$), while inequality (e) follows from the fact that $\sum_{i=1}^{n} a_i^2 \le \left(\sum_{i=1}^{n} a_i\right)^2$, 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 $\mathbb{E}\left\{\|\boldsymbol{\lambda}_2^n\|^2\right\}$ can be represented as

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

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

$$\begin{aligned}
&\mathbb{E}\{\Delta\} = \mathbb{E}\left\{\left\|\frac{\sum_{z=1}^{Z} N_z\left(\nabla F_z(\boldsymbol{w}_z^n)-\nabla F_z(\hat{\boldsymbol{w}}_z^n)\right)}{\sum_{z=1}^{Z} N_z}\right\|^2\right\} L^2\\
&\overset{(k)}{\le} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^{Z}\|N_z\|^2\right)\left(\sum_{z=1}^{Z}\|\nabla F_z(\boldsymbol{w}_z^n)-\nabla F_z(\hat{\boldsymbol{w}}_z^n)\|^2\right)}{\left\|\sum_{z=1}^{Z} N_z\right\|^2}\right\} L^2\\
&\overset{(j)}{\le} \mathbb{E}\left\{\frac{\left(\sum_{z=1}^{Z}\|N_z\|^2\right)\left(\sum_{z=1}^{Z}\|\boldsymbol{w}_z^n-\hat{\boldsymbol{w}}_z^n\|^2\right)}{\left\|\sum_{z=1}^{Z} N_z\right\|^2}\right\} L^2\\
&\overset{(l)}{\le} \frac{\left(\sum_{z=1}^{Z}\|N_z\|^2\right)\left(\sum_{z=1}^{Z}\mathbb{E}\left\{\|\boldsymbol{w}_z^n-\hat{\boldsymbol{w}}_z^n\|^2\right\}\right)}{\left\|\sum_{z=1}^{Z} N_z\right\|^2} L^2\\
&\overset{(l)}{\le} \frac{\left(\sum_{z=1}^{Z}\|N_z\|^2\right)\left(\sum_{z=1}^{Z} r_z^n D^2\right)}{\left\|\sum_{z=1}^{Z} N_z\right\|^2} L^2\\
&= \frac{\left(\sum_{z=1}^{Z}\|N_z\|^2\right)\left(\sum_{z=1}^{Z} r_z^n\right)}{\left\|\sum_{z=1}^{Z} N_z\right\|^2} L^2 D^2 \triangleq L^2 D^2 \Lambda_2^n,
\end{aligned} \tag{A.16}$$

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

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

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

$$\begin{aligned}
\mathbb{E}\left\{F\left(\boldsymbol{w}^{n+1}\right)\right\} \le{}& \mathbb{E}\left\{F(\boldsymbol{w}^n)\right\} - \frac{1}{L}\mathbb{E}\left\{\|\nabla F(\hat{\boldsymbol{w}}^n)\|^2\right\}\\
&+ \frac{ZA^2}{L} + LD^2\Lambda_2^n + \frac{\Lambda_1^n}{L}.
\end{aligned} \tag{A.18}$$

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

$$\begin{aligned}
\mathbb{E}\left\{\|\nabla F(\hat{\boldsymbol{w}}^n)\|^2\right\} \le{}& L\mathbb{E}\left\{F(\boldsymbol{w}^n)-F\left(\boldsymbol{w}^{n+1}\right)\right\} + UA^2\\
&+ L^2D^2\Lambda_2^n + \Lambda_1^n.
\end{aligned} \tag{A.19}$$

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

$$\begin{aligned}
&\frac{1}{\Omega+1}\sum_{n=0}^{\Omega}\mathbb{E}\left\{\|\nabla F(\hat{\boldsymbol{w}}^n)\|^2\right\} \le \frac{L}{\Omega+1}\mathbb{E}\left\{F\left(\boldsymbol{w}^0\right)-F\left(\boldsymbol{w}^*\right)\right\}\\
&+ UA^2 + \frac{L^2D^2}{\Omega+1}\sum_{n=0}^{\Omega}\Lambda_2^n + \frac{1}{\Omega+1}\sum_{n=0}^{\Omega}\Lambda_1^n.
\end{aligned} \tag{A.20}$$

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

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

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

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

where Λ^n is given by

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

This completes the proof. ∎

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.