Energy-Efficient Federated Learning: Integrating Model Pruning, Compressive Sensing, and Outage Compensation

Fangming Guan, *Student Member, IEEE,* Xiangwang Hou, *Student Member, IEEE,*

I. System Model

A. Transmission Model

In the uplink, each device uploads their local gradient employing orthogonal frequency division multiplexing (OFDM). In the t-th iteration, the upload rate of device u can be given by

$$R_u^t\left(p_u^t\right) = B_u^{\mathrm{UL}} \mathbb{E}_{h_u^t}\left(\log_2\left(1+\frac{p_u^t h_u^t}{I_u^t + B_u^{\mathrm{UL}} N_0}\right)\right), \quad (1)$$

where B_u^{UL} represents the uplink bandwidth of device u, h_u^t represents the channel gain, p_u^t represents the transmission power of device u, I_u^t represents communication interference, and N_0 represents the power spectral density of noise. The channel gain can be calculated by the following equation

$$h_u^t = \frac{\xi_u^t}{\left(d_u^t\right)^2}, \quad (2)$$

where ξ_u^t is the Rayleigh fading coefficient, d_u^t represents the distance between device u and the BS.

In practical wireless communication, transmission errors are unavoidable. Assuming that the gradient information from device u is uploaded as data packets, the probability of a transmission error for these packets can be represented as

$$q_u^t\left(p_u^t\right) = \mathbb{E}_{h_u^t}\left(1 - \exp\left(-\frac{\Upsilon\left(I_u^t + B_u^{\mathrm{UL}} N_0\right)}{p_u^t h_u^t}\right)\right), \quad (3)$$

where Υ is the waterfall threshold. We use a binary variable α_u^t indicate whether the local gradient of u is received successfully by the BS at the t-th iteration. α_u^t is given by

$$\alpha_u^t = \begin{cases} 1, & \text{if } 1 - q_u^t\left(p_u^t\right), \\ 0, & \text{if } q_u^t\left(p_u^t\right). \end{cases} \quad (4)$$

B. FL Model

Let's examine a scenario featuring a single base station (BS) and U local devices, denoted as $\mathcal{U} = \{1, 2, \ldots, u, \ldots, U\}$. Each user, indexed by u, possesses K_u data samples represented by $\mathcal{D}_u = \{(\boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k})\}_{k=1}^{K_u}$, where $K_u = |\mathcal{D}_u|$ is the number of training samples on user u. The union of all training datasets can be denoted as $\mathcal{D} = \bigcup_u \mathcal{D}_u$, and $K = \sum_{u=1}^{U} K_u$ is the number of total training samples. In $\mathcal{D}_u$, $(\boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k})$ represents the k-th data sample and its label. The primary aim of the training process is to minimize the global loss function, denoted as $F(w)$, associated with the global shared learning model, which is parameterized by $\boldsymbol{w} = [w^1, w^2, ..., w^N]$. Mathematically, this objective can be expressed as

$$\mathcal{P}1: \boldsymbol{w}^* = \arg\min F\left(\boldsymbol{w}\right), \quad (5)$$

where we define the global loss function as

$$F\left(\boldsymbol{w}\right) \triangleq \frac{\sum_{u=1}^{U} K_u \sum_{k=1}^{K_u} f\left(\boldsymbol{w}_u; \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right)}{K}. \quad (6)$$

Let the local learning model parameterized by $\boldsymbol{w}_u$, the loss function, evaluating training error on local dataset D_u, is defined as

$$F_u\left(\boldsymbol{w}_u\right) = \frac{1}{K_u}\sum_{k=1}^{K_u} f\left(\boldsymbol{w}_u; \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right), \quad \forall u \in \mathcal{U}. \quad (7)$$

To tackle problem $\mathcal{P}1$, the gradient descent algorithm operates at users in parallel. Firstly, each user updates its local gradient as follows

$$\boldsymbol{g}_u = \frac{1}{K_u}\sum_{k=1}^{K_u} \nabla f\left(\boldsymbol{w}_u; \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right), \quad u = 1, \ldots, U, \quad (8)$$

where $\nabla f\left(\boldsymbol{w}_u; \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right)$ is the gradient of $f\left(\boldsymbol{w}_u; \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right)$. Then each local device uploads their gradients to the base station (BS) and aggregates them according to the following procedure

$$\boldsymbol{g}^t = \frac{\sum_{u=1}^{U} K_u \alpha_u^t \boldsymbol{g}_u^t}{\sum_{u=1}^{U} K_u \alpha_u^t} \quad u = 1, \ldots, U. \quad (9)$$

After that, the global ML model update will be performed

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{g}^t, \quad (10)$$

where η denotes the learning rate. The global parameter $\boldsymbol{w}$ is then sent back to the local s for updating the shared model by

$$\boldsymbol{w}_u = \boldsymbol{w}^{t+1}. \quad (11)$$

Next, the FL iteratively repeat Eq. (8) - Eq. (11) to get an optimal FL model eventually.

X. Hou, J. Du, Y. Ren are with the Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China. (E-mail: hxw21@mails.tsinghua.edu.cn, {jundu, reny}@tsinghua.edu.cn.)

J. Wang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China. (Email: drwangjj@buaa.edu.cn)

C. Jiang is with the Tsinghua Space Center, Tsinghua University, Beijing, 100084, China. (E-mail: jchx@tsinghua.edu.cn).

C. 1-Bit Wireless FL

To facilitate the training of large-scale models capable of executing complex tasks within the FL framework, we integrate model pruning, compression sensing, and 1-bit quantization into the FL, using a fault compensation mechanism to enhance convergence speed.

1) Pruning: Before the t-th round of training, we define a mask $m_u^t \in \{0,1\}$ to prune the FL model. Let N denote the dimensionality of the parameter vector. For the parameter vector $\boldsymbol{w}_u^t$ of local device u, with a random gradient $\boldsymbol{g}(\boldsymbol{w}_u^t) \in \mathbb{R}^N$, to produce a sparse parameter vector, we apply the mask m_u^t to the parameter as

$$\tilde{\boldsymbol{w}}_u^t = m_u^t \bigodot \boldsymbol{w}_u^t, \tag{12}$$

where $\tilde{\boldsymbol{w}}_u^t$ represents the sparsified model, and $\bigodot$ denotes element-wise multiplication. The pruned model parameter $\tilde{\boldsymbol{w}}_u^t$ contains k_u^t non-zero components, with all other components set to zero. Subsequently, the model computes gradients according to Eq. (8), resulting in gradients that match the sparsity as the model parameters. Let ρ_u^t denote the pruning ratio of local device u in iteration t. The pruning ratio is defined as the ratio of the number of pruned parameters to the original number of parameters. When ρ_u^t = 0, it indicates that no pruning operation is performed in this iteration. For brevity, in writing, we represent the gradient computed from the pruned model as $\tilde{\boldsymbol{g}}_u^t(\tilde{\boldsymbol{w}}_u^t)$, abbreviated as $\tilde{\boldsymbol{g}}_u^t$. The pruned gradient information exhibits sufficient sparsity to meet basic CS requirements. We use z_u to denote the sparsity of $\tilde{\boldsymbol{g}}_u^t$, which determines our CS strategy.

2) Compression Sensing: To significantly reduce communication overhead in FL, we introduce CS to compress the local parameter information following sparsification. For the entire FL system, we introduce an measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{S \times N}$ $(S \ll N)$, which maps the high-dimensional $\tilde{\boldsymbol{g}}_u^t$ to a lower-dimensional space. This matrix is uniformly shared among all local devices and the BS. To enable signal reconstruction, the observation matrix must satisfy the restricted isometry property (RIP) condition, which requires it to be an independently and identically distributed Gaussian random matrix. The compressed local gradient is then represented as $\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t$.

3) 1-Bit Quantization: After applying CS to the local gradient, we then proceed with 1-bit quantization as

$$\bar{\boldsymbol{g}}_u^t = \boldsymbol{sign}(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t). \tag{13}$$

4) Reconstruction: After receiving the gradient information, the BS needs to use CS reconstruction algorithms to recover the original gradient information and then compute the global gradient using Eq. (9). Many effective algorithms can be applied to gradient reconstruction, such as. We use $CS^{-1}(\cdot)$ to represent the CS reconstruction operation on the gradient.

5) Accumulated Gradient Feedback: In each iteration, only a subset of devices participate. The gradients that fail to upload to BS are retained on the local device and are typically discarded. This could negatively impact model convergence or even result in training bias. To address this, we designed a compensation mechanism to retain local gradients that failed to transmit. The retained local gradients can be represented as

$$\boldsymbol{r}_u^t = \begin{cases} 0 & t = 1, \\ 0 & \alpha_u^{t-1} = 1, t \geq 2, \\ \rho\bar{\boldsymbol{g}}_u^{t-1} & \alpha_u^{t-1} = 0, t \geq 2. \end{cases} \tag{14}$$

When communication conditions allow, these retained gradients are uploaded to the BS along with the gradients generated in the current training round.Then the FL training process, as described by Eq. (9) and Eq. (10) will be reformulated as

$$\hat{\boldsymbol{g}}^t = CS^{-1}\left(\frac{\sum_{u=1}^{U} K_u \alpha_u^t \left(\bar{\boldsymbol{g}}_u^t + \boldsymbol{r}_u^t\right)}{\sum_{u=1}^{U} K_u \alpha_u^t}\right) \quad u = 1, \ldots, U, \tag{15}$$

and

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\hat{\boldsymbol{g}}^t. \tag{16}$$

II. Convergence Analysis and Problem Formulation

A. Basic Assumptions

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 with respect to* $\boldsymbol{w}$*, which can be given by*

$$\left\|\nabla F\left(\boldsymbol{w}^{t+1}\right)-\nabla F\left(\boldsymbol{w}^{t}\right)\right\| \leqslant L\left\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^{t}\right\|, \quad (17)$$

where L *is Lipschitz constant depending on* $F(\cdot)$*.*

- **Assumption 2:** $\nabla F(\boldsymbol{w})$ *is twice-continuously differentiable. Considering both* Assumption 1 *and* Assumption 2*, the following inequality can be established:*

$$\gamma \boldsymbol{I} \leqslant \nabla^2 F(\boldsymbol{w}) \leqslant L \boldsymbol{I}, \quad (18)$$

where $\mathbf{I}$ *is an identity matrix.*

- **Assumption 3:** *The second moments of local gradient and parameters are constrained by*

$$\left\|\nabla f\left(\boldsymbol{w}^{t}, \boldsymbol{x}_{u,k}, \boldsymbol{y}_{u,k}\right)\right\|^{2} \leqslant G^{2}, \quad (19)$$

and

$$\mathbb{E}\{\|\boldsymbol{w}\|^{2}\} \leqslant D^{2}. \quad (20)$$

- **Assumption 4:** *The stochastic gradients are unbiased, which can be represented as*

$$\mathbb{E}\{g(\boldsymbol{w})\} = \nabla F(\boldsymbol{w}). \quad (21)$$

APPENDIX A
PROOF OF LEMMA 2

The following two inequalities will be frequently used in the subsequent derivations.
Cauchy-Schwarz inequality

$$\sum_{k=1}^{n} {a_k}^2 \sum_{k=1}^{n} {b_k}^2 \geqslant \left(\sum_{k=1}^{n} a_k b_k\right)^2, \quad (A.1)$$

for $a_k \geqslant 0$

$$\sum_{k=1}^{n} {a_k}^2 \leqslant \left(\sum_{k=1}^{n} a_k\right)^2, \quad (A.2)$$

and the triangle inequality of the Euclidean norm

$$\|\boldsymbol{X}+\boldsymbol{Y}\| \leqslant \|\boldsymbol{X}\|+\|\boldsymbol{Y}\|. \quad (A.3)$$

Let $\hat{\boldsymbol{g}}^t$ denote the gradient reconstructed by compressed sensing. According to [**?**], the reconstruction error can be expressed as $\mathbb{E}\left\|\hat{\boldsymbol{g}}^{\boldsymbol{t}}-\tilde{\boldsymbol{g}}^{\boldsymbol{t}}\right\|^2$.Based on 1-bit quantization, feedback, and device dropout, the reconstruction error can be expressed as

$$\mathbb{E}\left\|\hat{\boldsymbol{g}}^{\boldsymbol{t}}-\tilde{\boldsymbol{g}}^{\boldsymbol{t}}\right\|^2 \leqslant \frac{C^2}{S}\mathbb{E}\left\|\bar{\boldsymbol{g}}^t+\boldsymbol{r}^t-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^t-\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^{t-1}|_{\alpha^t=0}\right)\right\|^2, \quad (A.4)$$

where C is the constant depending on the properties of $\boldsymbol{\Phi}$ and $0<\delta<1$ is a constant of the RIP condition. We can express the total quantization error as

$$\begin{aligned} &E\left\|\bar{\boldsymbol{g}}^t+\boldsymbol{r}^t-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^t-\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^{t-1}|_{\alpha^t=0}\right)\right\|^2 \\ &=\mathbb{E}\left\|\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left(\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}|_{\alpha_u^{t-1}=0}\right)}{\sum_{u=1}^{U} K_u\alpha_u^t}\right. \\ &\qquad\left.+\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left(\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t}\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t}\right)}{\sum_{u=1}^{U} K_u\alpha_u^t}\right\|^2. \end{aligned} \quad (A.5)$$

We will separately provide the upper bounds for the two aforementioned parts. First, we derive the upper bound of the gradient before CS and 1-bit quantization based on Assumption 3.

$$\begin{aligned} \mathbb{E}\|\boldsymbol{g}\|^2 &= \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla f\left(\boldsymbol{w}_u;\boldsymbol{x}_k,\boldsymbol{y}_k\right)\right\|^2 \\ &\overset{(c)}{\leqslant} \frac{1}{K^2}\mathbb{E}\left\{\sum_{k=1}^{K}\left\|\nabla f\left(\boldsymbol{w}_u;\boldsymbol{x}_k,\boldsymbol{y}_k\right)\right\|^2\right\} \overset{(d)}{\leqslant} \frac{G^2}{K}. \end{aligned} \quad (A.6)$$

where (c) comes from Eq. (A.3), and (d) arises from Assumption 3. Next, we derive the upper bound of the error caused by 1-bit quantization. Measurement matrix $\boldsymbol{\Phi}$ obeys ”restricted isometry hypothesis”, which means

$$(1-\delta)\|\boldsymbol{g}\|^2 \leqslant \|\boldsymbol{\Phi}\boldsymbol{g}\|^2 \leqslant (1+\delta)\|\boldsymbol{g}\|^2. \quad (A.7)$$

In the t-th iteration, the upper bound of the error caused by 1-bit quantization of each device’s gradient is given by

$$\begin{aligned} \mathbb{E}\left\|\boldsymbol{e}_{u,q}^t\right\|^2 &= \mathbb{E}\left\|\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right\|^2 \\ &\overset{(d)}{\leqslant} 2\mathbb{E}\left(\left\|\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right)\right\|^2+\left\|\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right\|^2\right) \\ &\leqslant 2S+2(\delta+1)\frac{G^2}{K}, \end{aligned} \quad (A.8)$$

where (d) is due to the fact that $(a-b)^2 \leqslant 2\left(a^2+b^2\right)$ and the error of the accumulated gradient is given by

$$\begin{aligned} \mathbb{E}\left\|\boldsymbol{e}_{qr}^t\right\|^2 &= \mathbb{E}\left\|\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}|_{\alpha_u^{t-1}=0}\right\|^2 \\ &\leqslant 2\mathbb{E}\left(\left\|\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right)\right\|^2+\left\|\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right\|^2|_{\alpha_u^{t-1}=0}\right) \\ &\leqslant 2S\sum_{u=1}^{U} q_u+2(\delta+1)\frac{G^2}{K}\sum_{u=1}^{U} q_u. \end{aligned} \quad (A.9)$$

The error after considering device dropout can be expressed as

$$\begin{aligned} &\mathbb{E}\left\|\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left(\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^t\right)}{\sum_{u=1}^{U} K_u\alpha_u^t}\right\|^2 \\ &\overset{(e)}{\leqslant} \mathbb{E}\left\{\left(\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left\|\boldsymbol{e}_{u,q}^t\right\|}{\sum_{u=1}^{U} K_u\alpha_u^t}\right)^2\right\} \leqslant max(\mathbb{E}\left\|\boldsymbol{e}_{u,q}^t\right\|^2) \\ &\qquad\qquad \leqslant 2S+2(\delta+1)\frac{G^2}{K}, \end{aligned} \quad (A.10)$$

and

$$\begin{aligned} &\mathbb{E}\left\|\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left(\boldsymbol{sign}\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}\right)-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}_u^{t-1}|_{\alpha_u^{t-1}=0}\right)}{\sum_{u=1}^{U} K_u\alpha_u^t}\right\|^2 \\ &\leqslant\mathbb{E}\left\{\left(\frac{\sum_{u=1}^{U} K_u\alpha_u^t\left\|\boldsymbol{e}_{qr}^t\right\|}{\sum_{u=1}^{U} K_u\alpha_u^t}\right)^2\right\} \leqslant max(\mathbb{E}\left\|\boldsymbol{e}_{qr}^t\right\|^2) \\ &\qquad\qquad 2S\sum_{u=1}^{U} q_u+2(\delta+1)\frac{G^2}{K}\sum_{u=1}^{U} q_u. \end{aligned} \quad (A.11)$$

Substituting Eq. (A.10) and Eq. (A.11) into Eq. (A.5), we can obtain

$$\begin{aligned}
&\mathbb{E}\left\|\hat{\boldsymbol{g}}^t-\tilde{\boldsymbol{g}}^t\right\|^2 \leqslant \frac{C^2}{S}\mathbb{E}\left\|\bar{\boldsymbol{g}}^t+\boldsymbol{r}^t-\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^t-\left(\boldsymbol{\Phi}\tilde{\boldsymbol{g}}^{t-1}|_{\alpha^t=0}\right)\right\|^2 \\
&\leqslant 6C^2+\frac{6(\delta+1)C^2G^2}{KS}+6C^2\sum_{u=1}^{U}q_u \\
&+\frac{6(\delta+1)C^2G^2\sum_{u=1}^{U}q_u}{KS}.
\end{aligned} \tag{A.12}$$

Appendix B
Proof of Theorem 1

When the BS obtains the information from local devices, it reconstructs the global gradient $\hat{\boldsymbol{g}}^t$ to update the FL model as

$$\boldsymbol{w}^{t+1}=\boldsymbol{w}^t-\eta\hat{\boldsymbol{g}}^t. \tag{B.1}$$

Performing a second-order Taylor expansion of $F\left(\boldsymbol{w}^{t+1}\right)$ at $\boldsymbol{w}^t$ and using Assumption 2, we can obtain

$$\begin{aligned}
F\left(\boldsymbol{w}^{t+1}\right) &\leqslant F\left(\boldsymbol{w}^t\right)+\left(\nabla F\left(\boldsymbol{w}^t\right)\right)^\top\left(\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\right) \\
&\quad+\frac{\nabla^2F\left(\boldsymbol{w}^t\right)}{2}\left\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\right\|^2 \\
&\leqslant F\left(\boldsymbol{w}^t\right)+\left(\nabla F\left(\boldsymbol{w}^t\right)\right)^\top\left(\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\right) \\
&\quad+\frac{L}{2}\left\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^t\right\|^2 \\
&\leqslant F\left(\boldsymbol{w}^t\right)-\eta\left(\nabla F\left(\boldsymbol{w}^t\right)\right)^\top\hat{\boldsymbol{g}}^t+\frac{L\eta^2}{2}\left\|\hat{\boldsymbol{g}}^t\right\|^2,
\end{aligned} \tag{B.2}$$

and

$$\begin{aligned}
&-\eta\left(\nabla F\left(\boldsymbol{w}^t\right)\right)^\top\hat{\boldsymbol{g}}^t \\
&=\frac{\eta}{2}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t\right\|^2-\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2-\left\|\hat{\boldsymbol{g}}^t\right\|^2\right\} \\
&\leqslant\frac{\eta}{2}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t\right\|^2-\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2\right\} \\
&=\frac{\eta}{2}\left\|\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t\right\|^2-\frac{\eta}{2}\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2.
\end{aligned} \tag{B.3}$$

So we obtain

$$\begin{aligned}
\mathbb{E}\left\{F\left(\boldsymbol{w}^{t+1}\right)\right\} \leqslant \mathbb{E}\Big\{&F\left(\boldsymbol{w}^t\right)+\frac{\eta}{2}\left\|\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t\right\|^2 \\
&-\frac{\eta}{2}\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2+\frac{L\eta^2}{2}\left\|\hat{\boldsymbol{g}}^t\right\|^2\Big\}.
\end{aligned} \tag{B.4}$$

For the sake of simplicity in writing, we introduce an auxiliary variable as

$$\boldsymbol{\lambda}^t=\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t, \tag{B.5}$$

which can be bounded by

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{\lambda}^t\right\|^2=\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\hat{\boldsymbol{g}}^t\right\|^2\right\} \\
&=\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\tilde{\boldsymbol{g}}^t(\tilde{\boldsymbol{w}}^t)-\left(\hat{\boldsymbol{g}}^t-\tilde{\boldsymbol{g}}^t\right)\right\|^2\right\} \\
&\leqslant 2\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\tilde{\boldsymbol{g}}^t(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}+2\mathbb{E}\left\{\left\|\hat{\boldsymbol{g}}^t-\tilde{\boldsymbol{g}}^t\right\|^2\right\}.
\end{aligned} \tag{B.6}$$

We have provided the upper bound for $\mathbb{E}\left\{\left\|\hat{\boldsymbol{g}}^t-\tilde{\boldsymbol{g}}^t\right\|^2\right\}$ in Appendix A, then we will present the upper bound for the first part. We introduce an auxiliary variable as

$$\nabla F(\boldsymbol{w}^t)'=\frac{\sum\limits_{u=1}^{U}K_u\alpha_u^t\nabla F_u\left(\boldsymbol{w}_u^t\right)}{\sum_{u=1}^{U}K_u\alpha_u^t}, \tag{B.7}$$

so we can obtain

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\} \\
&=\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F\left(\boldsymbol{w}^t\right)'+\nabla F\left(\boldsymbol{w}^t\right)'-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\} \\
&\leqslant\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F\left(\boldsymbol{w}^t\right)'\right\|^2+\left\|\nabla F\left(\boldsymbol{w}^t\right)'-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}.
\end{aligned} \tag{B.8}$$

Next, we will provide the upper bounds for the two parts above separately. Let U_1 represents the set of devices participating in the global iteration, and $|U_1|$ denotes the number of devices. Conversely, U_2 represents the set of devices that are not involved in the global iteration, and $|U_2|$ denotes the number of devices.

For the first part $\mathbb{E}\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F\left(\boldsymbol{w}^t\right)'\right\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\boldsymbol{w}^t)'\right\|^2\right\} \\
&=\mathbb{E}\left\{\left\|\frac{\sum\limits_{u=1}^{U}K_u\nabla F_u\left(\boldsymbol{w}_u^t\right)}{K}-\frac{\sum\limits_{u=1}^{U}K_u\alpha_u^t\nabla F_u\left(\boldsymbol{w}_u^t\right)}{\sum_{u=1}^{U}K_u\alpha_u^t}\right\|^2\right\},
\end{aligned} \tag{B.9}$$

based on Eq. (8), we can obtain

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\boldsymbol{w}^t)'\right\|^2\right\} \\
&\overset{(f)}{\leqslant}\mathbb{E}\left\{\left\|\sum_{u=1}^{U}\left(\frac{1}{K}-\frac{\alpha_u^t}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right)\cdot\right.\right. \\
&\qquad\qquad\left.\left.\sum_{k=1}^{K_u}\nabla f\left(\boldsymbol{w}_u;\boldsymbol{x}_{u,k},\boldsymbol{y}_{u,k}\right)\right\|^2\right\} \\
&\overset{(g)}{\leqslant}\mathbb{E}\left\{\sum_{u=1}^{U}\left\|\frac{1}{K}-\frac{\alpha_u^t}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right\|^2\cdot\right. \\
&\qquad\qquad\left.\sum_{u=1}^{U}\left\|\sum_{k=1}^{K_u}\nabla f\left(\boldsymbol{w}_u;\boldsymbol{x}_{u,k},\boldsymbol{y}_{u,k}\right)\right\|^2\right\}
\end{aligned} \tag{B.10}$$

where (f) arises from Eq. (8), (g) comes from Eq. (A.1). We use Assumption 3 to obtain

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\boldsymbol{w}^t)'\right\|^2\right\}\\
&\overset{(h)}{\leqslant}\mathbb{E}\left\{\sum_{u=1}^{U}\left\|\frac{1}{K}-\frac{\alpha_u^t}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right\|^2\sum_{u=1}^{U}K_uG^2\right\}\\
&\leqslant KG^2\mathbb{E}\left\{\sum_{u=1}^{U}\left(\frac{1}{K}-\frac{\alpha_u^t}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right)^2\right\}\\
&\leqslant KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right.\\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2}q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1}K_{u_1}}\right)\right\}^2,
\end{aligned}\tag{B.11}$$

For the second part

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)'-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}\\
&=\mathbb{E}\left\{\left\|\frac{\sum\limits_{u=1}^{U}K_u\alpha_u^t\left(\nabla F_u\left(\boldsymbol{w}_u^t\right)-\nabla F_u\left(\tilde{\boldsymbol{w}}_u^t\right)\right)}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right\|^2\right\}\\
&\overset{(i)}{\leqslant}\mathbb{E}\left\{\frac{\left(\sum_{u=1}^{U}\|K_u\alpha_u^t\|^2\right)\left(\sum_{u=1}^{U}\|\boldsymbol{w}_u^t-\tilde{\boldsymbol{w}}_u^t\|^2\right)}{\left\|\sum_{u=1}^{U}K_u\alpha_u^t\right\|^2}\right\}L^2\\
&=\frac{\left(\sum_{u=1}^{U}\|K_u\alpha_u^t\|^2\right)\left(\sum_{u=1}^{U}\mathbb{E}\left\{\|\boldsymbol{w}_u^t-\tilde{\boldsymbol{w}}_u^t\|^2\right\}\right)}{\left\|\sum_{u=1}^{U}K_u\alpha_u^t\right\|^2}L^2\\
&\overset{(j)}{\leqslant}\frac{\left(\sum_{u=1}^{U}\|K_u\alpha_u^t\|^2\right)\left(\sum_{u=1}^{U}\rho_u^tD^2\right)}{\left\|\sum_{u=1}^{U}K_u\alpha_u^t\right\|^2}L^2\\
&\overset{(k)}{\leqslant}\frac{\left(\sum_{u=1}^{U}\|K_u\alpha_u^t\|^2\right)\left(\sum_{u=1}^{U}\rho_u^t\right)}{\sum_{u=1}^{U}\|K_u\alpha_u^t\|^2}L^2D^2\triangleq L^2D^2\Gamma^t.
\end{aligned}\tag{B.12}$$

(i) is due to Assumption 1 and (k) is derived from Eq. (A.2). And we can obtain (j) from [2].

Based on the results of the two parts above, we can obtain

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}\\
&=\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F\left(\boldsymbol{w}^t\right)'+\nabla F\left(\boldsymbol{w}^t\right)'-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}\\
&\leqslant\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F\left(\boldsymbol{w}^t\right)'\right\|^2+\left\|\nabla F\left(\boldsymbol{w}^t\right)'-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}\\
&\leqslant L^2D^2\Gamma^t+KG^2\mathbb{E}\left\{\sum_{u=1}^{U}\left(\frac{1}{K}-\frac{\alpha_u^t}{\sum\limits_{u=1}^{U}K_u\alpha_u^t}\right)^2\right\}\\
&\leqslant L^2D^2\Gamma^t+KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right.\\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2}q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1}K_{u_1}}\right)\right\}^2.
\end{aligned}\tag{B.13}$$

$$\begin{aligned}
&E\left\|\nabla F\left(\boldsymbol{w}^t\right)-\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\|^2\\
&=E\|\nabla F(\boldsymbol{w}^t)\|^2+E\left\|\tilde{\boldsymbol{g}}^t(\tilde{\boldsymbol{w}}^t)\right\|^2\\
&\qquad-2E\left\{\left(\nabla F\left(\boldsymbol{w}^t\right)\right)^\top\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\}\\
&=E\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2+E\left\|\tilde{\boldsymbol{g}}^t(\tilde{\boldsymbol{w}}^t)\right\|^2\\
&\qquad-2E\left\{\nabla F\left(\tilde{\boldsymbol{w}}^t\right)\right\}^\top E\left\{\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\}\\
&\qquad-2\operatorname{Tr}\left(\operatorname{Cov}\left(\nabla F,\tilde{\boldsymbol{g}}^t\right)\right)\\
&\leqslant E\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2+E\left\|\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\|^2\\
&\qquad-2E\left\{\nabla F\left(\boldsymbol{w}^t\right)\right\}^\top E\left\{\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\}\\
&=E\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2+E\left\|\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\|^2\\
&\qquad-2E\left\{\nabla F\left(\boldsymbol{w}^t\right)\right\}^\top E\left\{\nabla F\left(\tilde{\boldsymbol{w}}^t\right)\right\}\\
&=E\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)-\nabla F(\tilde{\boldsymbol{w}}^t)\right\|^2\right\}-E\|\nabla F(\tilde{\boldsymbol{w}}^t)\|^2\\
&\qquad+E\left\|\tilde{\boldsymbol{g}}^t\left(\tilde{\boldsymbol{w}}^t\right)\right\|^2\\
&\leqslant L^2D^2\Gamma^t+\frac{G^2}{K}-E\|\nabla F(\tilde{\boldsymbol{w}}^t)\|^2\\
&\qquad+KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right.\\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2}q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1}K_{u_1}}\right)\right\}^2
\end{aligned}\tag{B.14}$$

Substituting Eq. (A.12) and Eq. (B.14)into Eq. (B.6), we can obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{\lambda}^t\right\|^2 \leqslant 2L^2D^2\Gamma^t + 2\frac{G^2}{K} - 2E\|\nabla F(\tilde{\boldsymbol{w}}^t)\|^2 + 12C^2\sum_{u=1}^{U} q_u^t \\
&+ 12C^2 + \frac{12(\delta+1)C^2G^2}{SK} + \frac{12(\delta+1)C^2G^2\sum_{u=1}^{U} q_u^t}{SK} \\
&+ 2KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 \\
&\leqslant 2L^2D^2\Gamma^t + 2\frac{G^2}{K} + 12C^2\sum_{u=1}^{U} q_u^t \\
&+ 12C^2 + \frac{12(\delta+1)C^2G^2}{SK} + \frac{12(\delta+1)C^2G^2\sum_{u=1}^{U} q_u^t}{SK} \\
&+ 2KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 .
\end{aligned}
\tag{B.15}
$$

Substituting Eq. (A.6) and Eq. (B.15) into Eq. (B.4), we can obtain

$$
\begin{aligned}
&\mathbb{E}\left\{F\left(\boldsymbol{w}^{t+1}\right)\right\} \leqslant \mathbb{E}\left\{F\left(\boldsymbol{w}^{t}\right)\right\} + \eta L^2D^2\Gamma^t + \frac{\eta G^2}{K} + 6\eta C^2 \\
&+ \frac{6\eta(\delta+1)C^2G^2}{SK} + 6\eta C^2\sum_{u=1}^{U} q_u^t - \frac{\eta}{2}\mathbb{E}\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2 \\
&+ \frac{6\eta(\delta+1)C^2G^2\sum_{u=1}^{U} q_u^t}{SK} + \frac{L\eta^2G^2}{2K} \\
&\qquad + \eta KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 ,
\end{aligned}
\tag{B.16}
$$

and we can easily get

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2 \leqslant \frac{2}{\eta}\mathbb{E}\left\{F\left(\boldsymbol{w}^t\right) - F\left(\boldsymbol{w}^{t+1}\right)\right\} + 2L^2D^2\Gamma^t \\
&+ \frac{2G^2}{K} + 12C^2 + \frac{12(\delta+1)C^2G^2}{SK} + 12C^2\sum_{u=1}^{U} q_u^t \\
&+ \frac{12(\delta+1)C^2G^2\sum_{u=1}^{U} q_u^t}{SK} + \frac{L\eta G^2}{K} \\
&\qquad + 2KG^2\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 .
\end{aligned}
\tag{B.17}
$$

We sum and average the above equation from $t=0$ to $\Omega-1$,

$$
\begin{aligned}
&\frac{1}{\Omega}\sum_{t=0}^{\Omega-1}\mathbb{E}\left\{\left\|\nabla F\left(\boldsymbol{w}^t\right)\right\|^2\right\} \leqslant \frac{2}{\eta\Omega}\mathbb{E}\left\{F\left(\boldsymbol{w}^0\right) - F\left(\boldsymbol{w}^*\right)\right\} \\
&+ \frac{2L^2D^2}{\Omega}\sum_{t=0}^{\Omega-1}\Gamma^t + \frac{12(\delta+1)C^2G^2}{SK} + \frac{12C^2}{\Omega-1}\sum_{t=1}^{\Omega-1}\sum_{u=1}^{U} q_u^t \\
&+ 12C^2 + \frac{2G^2}{K} + \frac{12(\delta+1)C^2G^2}{SK(\Omega-1)}\sum_{t=1}^{\Omega-1}\sum_{u=1}^{U} q_u^t + \frac{L\eta G^2}{K} \\
&\qquad + \frac{2KG^2}{\Omega}\sum_{t=0}^{\Omega-1}\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 .
\end{aligned}
\tag{B.18}
$$

Therefore we can obtain the upper bound of the Euclidean norm of the gradient as

$$
\begin{aligned}
&\frac{1}{\Omega}\mathbb{E}\left\{\left\|\nabla F\left(\tilde{\boldsymbol{w}}^t\right)\right\|^2\right\} \leqslant \frac{2}{\eta\Omega}\mathbb{E}\left\{F\left(\boldsymbol{w}^0\right) - F\left(\boldsymbol{w}^*\right)\right\} \\
&+ \frac{2L^2D^2}{\Omega}\sum_{t=0}^{\Omega-1}\Gamma^t + \frac{12(\delta+1)C^2G^2}{SK} + \frac{12C^2}{\Omega-1}\sum_{t=1}^{\Omega}\sum_{u=1}^{U} q_u^t \\
&+ 12C^2 + \frac{2G^2}{K} + \frac{12(\delta+1)C^2G^2}{SK\Omega}\sum_{t=1}^{\Omega}\sum_{u=1}^{U} q_u^t + \frac{L\eta G^2}{K} \\
&\qquad + \frac{2KG^2}{\Omega}\sum_{t=0}^{\Omega-1}\sum_{i=1}^{U}\sum_{\substack{|U_1|=i\\|U_2|=U-i}}\left\{\frac{1}{K}-\right. \\
&\left.\left(\prod_{u_1\in U_1}\left(1-q_{u_1}^t\right)\prod_{u_2\in U_2} q_{u_2}^t\frac{1}{\sum\limits_{u_1\in U_1} K_{u_1}}\right)\right\}^2 ,
\end{aligned}
\tag{B.19}
$$

where $\boldsymbol{w}^*$ represents the final model.