Appendix A  Adjusting the Clipping Value

Considering the sparsification process, we have $g_i^t(D_{i,m}) = g_i^t(D_{i,m}) \odot m_i^t$, where $D_{i,m}$ is the $m$-th sample of the $i$-th client, $\odot$ represents the element-wise product process, $m_i^t$ is a binary mask vector and its element $\forall m_{i,k} \in \{0, 1\}$, $k \in \{1, \ldots, K\}$ and $K = |m_i^t|$. Because probabilities of $m_{i,k} = 0$ and $m_{i,k} = 1$ are equal to $s_i^t$ and $(1 - s_i^t)$, respectively, we have

\[
E[\|g_i^t(D_{i,m}) \odot m_i^t\|^2] = \sum_{k=1}^{K} \frac{k(K-1)}{K} (s_i^t)^{k-1} (1 - s_i^t)^{K-k} \sum_{m_{i,k}} (g_i^t(D_{i,m}))^2 = s_i^t \|g_i^t(D_{i,m})\|^2.
\]

Therefore, we can obtain $E[\|g_i^t(D_{i,m}) \odot m_i^t\|] \leq \sqrt{s_i^t} \|g_i^t(D_{i,m})\|$. This completes the proof.

Appendix B  Convergence Analysis

First, we define $1^i = \sum_{j=0}^{N} a_{i,j}^t$ to denote whether the $i$-th client has been allocated to an available channel. We can note that

\[
w^{t+1} - w^t = \sum_{i \in U} \sum_{t=0}^{T-1} p_{i}^t \nabla F_i(w_i^t) \odot m_i^t + n_i^{t+1} \odot m_i^t.
\]

Using the $L$-Lipschitz smoothness, we can obtain

\[
E[F(w^{t+1}) - F(w^t)] \leq E[\|\nabla F(w^t), \nabla F(w^{t+1}) - \nabla F(w^t)] + \frac{L}{2} E[\|w^{t+1} - w^t\|^2]
\]

\[
= -E \left[ \left( \sum_{i \in U} \sum_{t=0}^{T-1} p_{i}^t \nabla F_i(w_i^t) \odot m_i^t + n_i^{t+1} \odot m_i^t \right) \right] + \frac{\eta^2 L}{2} E \left[ \sum_{i \in U} \sum_{t=0}^{T-1} (\nabla F_i(w_i^t) \odot m_i^t + n_i^{t+1} \odot m_i^t) \right]
\]

\[
E_1 + \frac{\eta^2 L}{2} E_2
\]

Then, we can rewrite $E_1$ as

\[
E_1 = -\eta \sum_{i \in U} \sum_{t=0}^{T-1} p_{i}^t \nabla F_i(w_i^t), E[\nabla F_i(w_i^t) \odot m_i^t]) - \eta \sum_{i \in U} \sum_{t=0}^{T-1} p_{i}^t \nabla F_i(w_i^t), E[n_i^{t+1} \odot m_i^t])
\]

Because $E[n_i^{t+1}] = 0$ and $\langle x, y \rangle = \frac{1}{2}(\|y\|^2 + \|x\|^2 - \|x - y\|^2)$, we have

\[
E_1 = \frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w^t)\|^2 - \frac{\eta}{2} \sum_{i \in U} \sum_{t=0}^{T-1} E[\|\nabla F_i(w_i^t) \odot m_i^t\|^2] + \frac{\eta}{2} \sum_{i \in U} \sum_{t=0}^{T-1} E[\|\nabla F(w^t) - \nabla F_i(w_i^t) \odot m_i^t\|^2].
\]

\[
(\text{B4})
\]
Further, we have
\begin{align}
E_1 &= -\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w_t)\|^2 - \frac{\eta}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}}) \| \cdot m_i^t] \right) \\
&\quad + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. 
\end{align}

Due to Jensen’s inequality and (A1), we obtain
\begin{align}
E_1 &\leq -\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w_t)\|^2 - \frac{\eta}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. 
\end{align}

Then, we can bound $E_{11}$ as
\begin{align}
E_{11} &= E[\|w_t - (w_{t-1}^{i^{\ell-1}} - \eta \nabla F_i(w_{t-1}^{i^{\ell-1}}) \cdot m_i^t + n_i^t \cdot m_i^t)] = \eta^2 E[\left( \sum_{t=0}^{T-1} (\nabla F_i(w_{t-1}^{i^{\ell-1}}) \cdot m_i^t + n_i^t \cdot m_i^t) \right]^2] \\
&\leq \eta^2 \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_{t-1}^{i^{\ell-1}}) \cdot m_i^t + n_i^t \cdot m_i^t]^2 \right) = \eta^2 \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_{t^{i^{\ell-1}}})\|^2] + \eta^2 \left( \sum_{t=0}^{T-1} E[\|n_i^t \|^2] \right) \right).
\end{align}

Hence, we can obtain
\begin{align}
E_1 &\leq -\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w_t)\|^2 + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. \\
&\quad + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. 
\end{align}

Combining $E_1$ and $E_{11}$, we can obtain
\begin{align}
E[F(w_{t+1}) - F(w_t)] &\leq -\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w_t)\|^2 + \frac{\eta}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + \frac{3\eta T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. 
\end{align}

To ensure the training performance, we will select a proper DP noise variance to have $E[\|n_i^t \|^2] = E[\|n_i^t \|^2] \leq \Theta$. Due to the bounded gradient, by setting $\eta \|r\| < 1$ and $\eta^2 \|L\|^2 < 1$, we obtain
\begin{align}
E[F(w_{t+1}) - F(w_t)] &\leq -\frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla F(w_t)\|^2 + \frac{3\eta^2 T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + 3\eta \left( \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. 
\end{align}

Rearranging and summing $t$ from $0$ to $T - 1$, we have
\begin{align}
\frac{1}{T} \sum_{t=0}^{T-1} E[\|\nabla F(w_t)\|^2] &\leq \frac{2(F(w_0) - F(w_T))}{\eta T} + 3\varepsilon + \frac{3\eta^2 T^2}{2} \sum_{i \in \mathcal{U}} p_i^t \left( \sum_{t=0}^{T-1} E[\|\nabla F_i(w_t^{i^{\ell-1}})\|^2] + 3\eta \left( \sum_{i \in \mathcal{U}} p_i^t \varepsilon_i. \right) \right)
\end{align}

This completes the proof. \(\square\)

### Appendix C Solution of the Optimal Gradient Sparsification Rate
To obtain the optimal gradient sparsification rate, we first derive the relation between $s_i$ and $V^t$. Hence, we first consider the condition $Q^{drev} > 0$ and have
\begin{align}
V^t &= \sum_{i \in \mathcal{U}, j \in \mathcal{N}} (Q_i^{f, a} - \lambda p_i^t) a_{i,j} + Q^{drev} (d^t - d^{drev}) - \sum_{i \in \mathcal{U}, j \in \mathcal{N}} \beta_i = \sum_{i \in \mathcal{U}, j \in \mathcal{N}} (Q_i^{f, a} - \lambda p_i^t) a_{i,j} \\
&\quad + Q^{drev}_{\text{max}} \sum_{i \in \mathcal{N}} s_i \left( \frac{Z_i^{drev} a_{i,j}}{B \log_2 (1 + \frac{Z_i^{drev} a_{i,j}}{\sigma^2})} + d_i^{drev} + \frac{\tau}{f_i} \right) - Q^{drev} d^{drev} \sum_{i \in \mathcal{N}} \beta_i.
\end{align}
Due to the maximizing process, this problem can be divided into $N$ subproblems based on the client with the maximum delay. First, let us discuss the condition that the delay of the client owning the $j$-th channel is the maximum one among all clients. We assume the $j$-th channel is allocated to the $i$-th client and its delay is the maximum one. Thus, we can obtain

\[
d_{i,j} = \frac{Z_{j}^{i} s_{i}^{t}}{B \log_{2} (1 + \frac{p_{i,j}^{t} h_{i,j}^{t}}{\sigma})} + d_{i,do} \geq \max_{j \in U / i} \left\{ \sum_{j' \in \mathcal{N}} a_{i,j'}^{t} \left( \frac{Z_{j}^{i} s_{i}^{t}}{B \log_{2} (1 + \frac{p_{i,j'}^{t} h_{i,j'}^{t}}{\sigma})} + d_{i,do} + \frac{\tau |D| \Phi_{j'}}{f_{j'}} \right) \right\}.
\]

(C2)

From the above inequality, we have

\[
s_{i}^{t} \geq s_{i}^{t,\min} = \frac{B \log_{2} (1 + \frac{p_{i,j}^{t} h_{i,j}^{t}}{\sigma})}{\max_{j \in U / i} \left\{ \sum_{j' \in \mathcal{N}} a_{i,j'}^{t} \left( \frac{Z_{j}^{i} s_{i}^{t}}{B \log_{2} (1 + \frac{p_{i,j'}^{t} h_{i,j'}^{t}}{\sigma})} + d_{i,do} + \frac{\tau |D| \Phi_{j'}}{f_{j'}} \right) \right\}}.
\]

(C3)

If $s_{i}^{t,\min} > 1$, there is no solution to this subproblem. Otherwise, we can derive the first derivative of $V^{i}$ with respect to $s_{i}^{t}$ of the $j$-th subproblem as follows:

\[
\frac{\partial V^{i}}{s_{i}^{t}} = -\lambda + \frac{Z_{j}^{i} Q_{i,j}^{t,de}}{B \log_{2} (1 + \frac{p_{i,j}^{t} h_{i,j}^{t}}{\sigma})}.
\]

(C4)

If $\frac{Z_{j}^{i} Q_{i,j}^{t,de}}{B \log_{2} (1 + \frac{p_{i,j}^{t} h_{i,j}^{t}}{\sigma})} \leq \lambda$, it can be found that as the value of $s_{i}^{t}$ increases, the objective $V^{i}$ decreases. Hence, we have $s_{i}^{t,*} = 1$. For other fast clients, i.e., $i' \in U / i$, we have

\[
s_{i'}^{t} \leq s_{i'}^{t,\max} = \frac{B \log_{2} (1 + \frac{p_{i',j}^{t} h_{i',j}^{t}}{\sigma})}{\sum_{j' = 1}^{N} a_{i',j'}^{t} B \log_{2} (1 + \frac{p_{i',j}^{t} h_{i',j}^{t}}{\sigma})}.
\]

(C5)

We can also derive the first derivative of $V^{i}$ with respect to $s_{i'}^{t}$ as $\frac{\partial V^{i}}{s_{i'}^{t}} = -p_{i'}^{t} \lambda$. Therefore, we have $s_{i'}^{t,*} = \min \{ s_{i'}^{t,\max}, 1 \}$.

If $\frac{Z_{j}^{i} Q_{i,j}^{t,de}}{B \log_{2} (1 + \frac{p_{i,j}^{t} h_{i,j}^{t}}{\sigma})} > \lambda$, we can find that as the value of $s_{i}^{t}$ increases, the objective $V^{i}$ increases. Therefore, the system need to select a small gradient sparsification rate $s_{i}^{t}$ in $[a_{i}^{th}, 1]$ for the $i$-th client. However, for the $i'$-th client, i.e., $i' \in U / i$, we want to select a large gradient sparsification rate in $[s_{i'}^{th}, \min \{ 1, s_{i'}^{t,\max} \}$ because the first derivative of $V^{i}$ with respect to $s_{i'}^{t}$ is negative. We can decrease the $s_{i}^{t}$ from 1 and then $s_{i'}^{t,\max}$ may be selected. Therefore, the first derivative of $V^{i}$ with respect to $s_{i}^{t}$ should be modified because $s_{i'}^{t,\max}$ is related to $s_{i}^{t}$. When the first derivative of $V^{i}$ with respect to $s_{i}^{t}$ become negative, let us stop decreasing the value of $s_{i}^{t}$. We can note that this way can obtain the optimal $s_{i}^{t,*}$ and $s_{i'}^{t,*} = \min \{ s_{i'}^{t,\max}, 1 \}$.

Other subproblems can be addressed using the same method. Overall, after addressing all $N$ subproblems, the optimal solution can be obtained as the final output. This completes the proof.

\[\square\]

Appendix D Feasibility Analysis

We first introduce the Lyapunov function $\Gamma(Q^{t}) = \frac{1}{2}(Q^{t,de})^{2} + \frac{1}{2} \sum_{i \in U}(Q^{t,fa})^{2}$, in which the drift from one communication round can be given as

\[
\Gamma(Q^{t+1}) - \Gamma(Q^{t}) = \frac{1}{2} \left( \max \left\{ Q^{t,de + d^{Avg} - d^{*}}, 0 \right\} \right)^{2} - \frac{1}{2} \left( Q^{t,de} \right)^{2} + \frac{1}{2} \sum_{i \in U} \left( \max \{ Q_{i}^{t-1,fa + 1 - \beta_{i}}, 0 \} \right)^{2} - \frac{1}{2} \sum_{i \in U} (Q_{i}^{t,fa})^{2}
\]

\[
\leq \frac{1}{2} \left( d^{Avg} - d^{*} \right)^{2} + Q^{t,de + d^{Avg} - d^{*}} + \frac{1}{2} \sum_{i \in U} (1 - \beta_{i})^{2} + \sum_{i \in U} Q_{i}^{t,fa} (1 - \beta_{i}).
\]

(D1)

Because

\[
V^{i}(P^{t}, s^{t}, a^{t}) = -\lambda \sum_{i \in U} p_{i}^{t} s_{i}^{t} + \sum_{i \in U} Q_{i}^{t,fa} (1 - \beta_{i}) + Q^{t,de} (d^{Avg} - d^{*})
\]

we have

\[
V^{i}(P^{t}, s^{t}, a^{t}) \leq -\lambda \sum_{i \in U} p_{i}^{t} s_{i}^{t} + \frac{1}{2} \left( d^{Avg} - d^{*} \right)^{2} + Q^{t,de} (d^{Avg} - d^{*}) + \frac{1}{2} \sum_{i \in U} (1 - \beta_{i})^{2} + \sum_{i \in U} Q_{i}^{t,fa} (1 - \beta_{i}).
\]

(D3)

Due to $s_{i}^{t}, p_{i}^{t} \leq 1, a_{i,j}^{t} \in \{0, 1\}$, $\beta_{i} \leq \frac{1}{2}$, $d^{*} = \max_{i \in U} \sum_{j \in N} s_{i,j}^{t}$, and $d_{i,j}^{t} = d_{i,j}^{do} + d_{i,j}^{lo} + d_{i,j}^{up}$, $\forall i \in U, j \in N$, we have

\[
\mathbb{E}[V^{i}(P^{t}, s^{t}, a^{t})] \leq C_{1} + Q^{t,de} \mathbb{E} [d^{Avg} - d^{*}] + \sum_{i \in U} Q_{i}^{t,fa} \mathbb{E} [1 - \beta_{i}].
\]

(D4)
Based on Theorem 4.5 in [1] and Lemma 1 in [2], existing $\zeta^{\text{opt}} > 0$, we can obtain the following inequality:

$$E\left[V_t(P_t, s_t, a_t')\right] \leq C_1 + \zeta^{\text{opt}}. \quad \text{(D5)}$$

By summing this equation over $t = 0, 1, \ldots, T$, we obtain

$$\lim_{T \to \infty} \sup \frac{1}{T} \sum_{t=0}^{T-1} E\left[V_t(P_t, s_t, a_t')\right] \leq C_1 + \zeta^{\text{opt}} < \infty. \quad \text{(D6)}$$

This completes the proof. \qed

References

1. Neely M. J. Stochastic network optimization with application to communication and queueing systems. Synthesis Lectures on Communication Networks, 2010